Recent Research Reviews Journal (ISSN: 2583-7079)
https://irojournals.com/rrrj

From Barriers to Boosters: Optimizing Open Government Data through Publishing Guidelines and Data Protection Strategies

Khadidja Bouchelouche[1], AbdessamedRéda Ghomari[2], Leila Zemmouchi-Ghomari[3]

[1]LMCS Laboratory, Ecole Nationale Supérieure d'Informatique, ESI, Algiers, Algeria

[2]Ecole Nationale Supérieure des Technologies Avancées, ENSTA, Algiers, Algeria

E-mail: [1]k_bouchelouche@esi.dz, [2]a_ghomari@esi.dz, [3]leila.ghomari@ensta.edu.dz

Abstract

Open Government Data (OGD) is a global endeavor, a collaborative effort between governments worldwide to share datasets that encapsulate a wide spectrum of government activities, from environmental issues like pollution and climate to social aspects like education and childcare, and urban concerns like traffic and congestion, and healthcare statistics. As governments, being among the largest producers and collectors of data, are making OGD available online in diverse formats, primarily Word, PDF, or Excel, they are contributing significantly to this global initiative. The OGD initiative holds immense potential to revolutionize the way we access and use government data. Its primary objective is to enhance the discoverability, accessibility, and availability of data in alternative and preferably machine-readable formats. This, in turn, empowers a diverse set of stakeholders to develop innovative data applications under licensing schemes that permit unrestricted reuse. Despite these promising aspects, challenges such as data heterogeneity, data protection, data quality, and data provenance issues persist. This study aims to analyze and categorize these challenges and obstacles that hinder the OGD initiative from realizing its full potential, with a particular emphasis on data protection and security concerns for data providers.

Keywords: Open Government Data (OGD), Data Providers, OGD Publication, Data Protection issues, Data Publication Guidelines

1. Introduction

Open Government Data (OGD) is a pivotal international collaboration between the United States, the United Kingdom, France, and Singapore governments. It serves as a platform to share machine-readable datasets that cover a wide range of government activities (Valli Buttow & Weerts, 2022). These datasets, either produced by governments or under their control, encompass diverse information, including pollution/climate, education/childcare, and traffic/congestion statistics (Attard et al., 2015).

Governments worldwide are among the biggest producers and collectors of data. OGD (https://www.data.gov/) is available online in various formats, primarily Word, PDF, or Excel, concerning multiple fields (economy, agriculture, education, health, etc.) (Attard et al., 2015).

(Gottfried et al., 2021) The fundamental premise is that data will become more discoverable, accessible, and available in alternative formats; furthermore, diverse stakeholders will devise innovative data applications under licensing schemes that permit unrestricted reuse. Thus, several publications in the literature offer guidelines for publishing data on the Web. The basis of most of these guidelines is the eight principles of OGD.

A large number of applications have been developed that exploit the OGD (https://www.data.gov/applications) in different countries and offer many services to people wishing to obtain practical information concerning, for example, the distribution of job applications by sector of activity and by region of a country or electricity consumption according to the type of household appliance used by time slot of the day in another country or even more generally the foods to avoid or to advocate in the case of this or that disease.

The government must do more than publish OGD online to achieve the above-mentioned purposes. For this reason, several procedures must be followed (Attard et al., 2015).

It is imperative to respect publishing guidelines, principles, and standards. Also, respect the OGD standard process that allows preparing data for publishing, using the published data, and maintaining its sustainability (Bouchelouche et al., 2021). Moreover, data quality is a crucial axis that influences several aspects (López Reyes and Magnussen, 2022).

Despite the efforts of several commitments—including those about the publication of OGD (Šlibarand Mu, 2022), the exploitation of OGD (Spalević et al., 2023), and data quality (Bouchelouche et al., 2022) and (Quarati, 2023), among others, have failed to stimulate the

OGD initiative. Moreover, this is linked to different aspects for several reasons, considering the three fundamental dimensions (Data providers issues, Stakeholder issues, and data quality issues) (Attard et al., 2015).

Looking at the current literature, many works treated the process of publishing OGD for improvement based on a set of best practices and guidelines for OGD publication and exploitation (Solar et al., 2012); (Liu et al., 2011). Since the OGD initiative is not reaching its full potential yet, the new works still attempt to cover all possible challenges or define OGD challenges for specific countries, which leads to discovering the same kind of challenges (Attard et al., 2015). There is still a significant gap in the absence of technical commitments regarding data protection and security for data providers' dimensions (Attard et al., 2015).

Using technical guidelines, we intend to analyze and categorize the issues and obstacles that prevent the OGD initiative from reaching its maximum potential, with a particular emphasis on data protection and security concerns for data providers.

The organization of this paper is as follows: Section II is dedicated to analyzing the process of OGD publication to discover the existing guidelines and extract the main challenges faced by the OGD initiative. Section III will focus on OGD challenges to classify the founded challenges. Section IV provides technical guidelines for data providers' protection and security based on the challenges in the current literature. Section V presents the conclusion and perspectives.

2. OGD Publishing Guidelines Analysis

Publishing data is regarded as a fundamental part of OGD initiatives. By disseminating published data for public use, data publishing facilitates the achievement of OG's primary objective: the reuse, distribution, and application of published data (Attard et al., 2015).

(Solar et al., 2012), (Liu et al. 2011) have proposed a set of good practices for data publishing on the Web. The recommendations are mainly derived from the Eight Open Government Data Principles, formulated by a working group led by Carl Malamud on December 8th, 2007, in Sebastopol.

1. Ensure that public data is fully accessible and not subject to restrictions based on privilege, security, or privacy concerns;
2. Primary: The data from the source must be provided in its original form without any alterations or combining with other data.
3. Prompt: Once the data is generated, it should be promptly released to the public to maintain its value.
4. Accessible: The data is readily available to all potential consumers without any restrictions on its use;
5. Data should be supplied in a structured format for automated processing.
6. Non-discriminatory: Users can access the provided data without registering.
7. Non-Proprietary: The format of the disclosed data is not controlled exclusively by a single party;
8. License-Free: The data can be used without restriction by confidential commercial regulations, trademarks, patents, or copyrights.

These eight principles facilitate stakeholders' effective public data utilization, promoting a successful Open Government Data (OGD) initiative. In addition, they have considered a roadmap for publishing data (Attard et al., 2015).

Berners-Lee's Five Star Scheme (2010) provides a technical guide for Open Government Data (OGD) standards. It outlines the principles for each level of stars: open license, structured format, non-proprietary, structured, and machine-readable format, using open standards from W3C, and linking published data with existing data. The scheme requires only common principles for one, two, and three stars and requires open standards from W3C for four stars. For five stars, the scheme requires open standards from W3C, links between good OGD, and allowing stakeholders to consume the best potential of published data. The scheme also serves as a roadmap for data publishers to provide context for their published and existing data. The principles in Table 1 represent the typical differences between the Eight Open Government Data Principles and the Five Star Scheme. Table 1. Illustrates the principles adhere to the Eight Open Government Data Principles and the Five Star Scheme.

Table 1. Eight Open Government Data Principles and the Five Star Scheme.

	Complete	Primary	Timely	Accessible	Machine Processable	Non-Discriminatory	Non-Proprietary	License-Free
*				✓				✓
**				✓	✓			✓
***				✓	✓		✓	✓
****				✓	✓		✓	✓
*****				✓	✓		✓	✓

The W3C e-Government Interest group has released recommendations for publishing Open Government Data (OGD) to allow its public utilization. These rules recommend using patterned, persistent, and discoverable URLs/URIs to facilitate the easy consumption and location of the material. Documentation enhances data clarity, facilitating its discoverability. Linking data facilitates the establishment of connections between various datasets and associated documentation, enhancing contextual understanding. Dataset versioning enables users to reference and connect to previous and current versions, allowing new datasets to reference the original version. The interfaces should possess the capability of being interpreted by machines and be easily understandable by humans, hence enabling external entities to develop their interfaces. The data should be disseminated independently from the interface. Standardized names/URIs for all government objects are crucial for ensuring legitimacy, improving metadata, and facilitating discoverability.

In addition to the above, choosing which data to publish is significant and discussed by the W3C[1] e-Government Interest Group, as well as the suitable format for data publishing and its restrictions (Attard et al., 2015). After tackling privacy and security issues, the data must conform to applicable regulations and laws. To allow easy manipulation of the published data, it must be published in its raw form and serialized in RDF and XML format. The established open standards are also recommended by (Attard et al., 2015). Lastly, clear documentation on regulatory restrictions or legalities for using the published data should be presented.

3. Classification of OGD Initiative Challenges

To cover the possible existing challenges of OGD initiatives, we analyzed the current works treating the challenges (Choenni et al., 2022); (Rashideh et al., 2022); (Attard et al., 2015).

This study investigates the obstacles impeding the OGD effort's maximum potential. These obstacles can be classified into three categories: compromising stakeholder satisfaction with OGD, obstructing data openness, and deterring companies from participating in the initiative. The purpose of these challenges is to enhance the overall efficiency and effectiveness of the effort.

When dealing with the current literature, it is evident that interest in the current works that aim to go beyond the current challenges, where the authors focused on the two dimensions that address stakeholder satisfaction and data quality according to opening standards (Bouchelouche et al., 2022), (Quarati, 2023), (Attard et al., 2015).

Thus, this work focuses on the third-dimensional category, corresponding to the challenges discouraging entities from joining the OGD initiative.

Noting that even the new works regarding OGD challenges (Choenni et al., 2022), (Rashideh et al., 2022) still attempt to cover all possible challenges since the OGD initiative is not yet reaching its full potential or defining OGD challenges for specific countries which lead to discovering the same kind of challenges (Choenni et al., 2022); (Rashideh et al., 2022). Therefore, Table 2 represents the extracted challenges, classified into three main categories:

- End-user satisfaction: Challenges that affect stakeholders' satisfaction with using OGD;
- Quality of data available as public data: Challenges hindering data from being genuinely open;
- Publication of data by government entities: Challenges discouraging entities from joining the OGD initiative. The Table.2 shows the classification of OGD challenges.

Table 2. OGD Challenges Classification

Category	Challenges	Description
End-user satisfaction	Usability	Creating value requires stakeholders' active involvement in the data consumption process. Disclosure of data without any intention of exploitation is devoid of any purpose. The value-creation potential of the evaluated initiative is directly proportional to the usability of OGD (Attard et al., 2015).
	Timeliness	Timeliness enables measuring if the published data or metadata is up to date. For example, specific data might only be valuable if made openly available shortly after its development (Attard et al., 2015); (Choenni et al., 2022).
	Data Accuracy	The metadata quality directly impacts datasets' discoverability, as it facilitates the stakeholders' ability to locate the dataset(Attard et al., 2015); (Choenni et al., 2022).
	Data formats	The main objective of utilizing portals to initiate and disseminate data is facilitating its utilization, redistribution, and repurposing. To achieve economic progress, data providers (governmental entities) must consider the needs and demands of the data end-users. This will encompass the designated forms suitable for the broadest range of consumers. The W3C recommends utilizing open tools and standards, such as RDF and XML, as a framework for disseminating information. One potential resolution is to mandate that government organizations release their data in formats not owned by any particular company and that machines can quickly process them (Attard et al., 2015).
	Data ambiguity	Understanding and linking the published data requires extra effort to overcome semantic ambiguity. According to (Valli Buttow & Weerts, 2022), even the availability of the data in a machine-readable format is only beneficial if it is easily understandable; background knowledge regarding the subject may reach this. An easy, adequate solution for this ambiguity is using descriptive titles while publishing data or providing keys to code names.
	Data discoverability	This is a significant challenge because it is linked to metadata quality, which describes data that is only sometimes accurate or complete.

	Data provenance	The challenge regarding provenance appears when the data moves vertically, where several parallel entities collect it. In this case, the data could have duplicates, which we call overlapping scope, in addition to the modified and new data (López Reyes and Magnussen, 2022). Consequently, the way of manipulating and changing the data during the publishing process belongs to provenance. Public institutions are concerned about potential liability for any harm caused by using the data they supply due to its outdated, misinterpreted, or inaccurate nature. To alleviate this concern, several public organizations enforce limitations on the utilization of the provided data or refrain from making their data publicly available, hence leading to data that is not accessible. A practical approach to address this issue is facilitating social engagement regarding the relevant data (López Reyes & Magnussen, 2022). One such approach is to utilize named graphs.
Quality of data available as public data	Conflicting regulations	This challenge involves how OGD projects interact, leading to the potential for questionable utilization of critical data. This matter pertains to individuals or entities who generate and utilize data, enabling it to be accessible and transparent, even if it is subject to a well-defined legal structure (Zine et al., 2022).
	Privacy and data protection	Pursuing accountability, transparency, and open data is fundamentally at odds with the right to privacy and data preservation. The amalgamation of multiple published datasets may yield individualized nature data. Therefore, additional research is required to resolve this dispute without limiting the degree of data accessibility (Zine et al., 2022).
	Copyright and licensing	The licensing issue has two aspects. The first one is the incompatibility of licenses. Data should be published openly, enabling the unrestricted use, reuse, and distribution. The second is copyright inconsistencies from data sharing because of unclear dataset ownership. This prevents data publishing (Attard et al., 2015). There exist four open data licenses (Bouchelouche et al., 2022): • Creative Commons (CC) Licences[2]. • Creative Commons Zero (CC0)[3]. • Open Data Commons[4]. • The Open Government Licence.
		Companies who invest in creating their data will consider opening up their data as unfair competition because it can create new competitors

	Competition	who did not invest anything to have free open data. For this, management mechanisms must be applied to guarantee that private companies' finances do not suffer and that they can open their data (Zine et al., 2022).
Publication of data by government entities	Awareness	Even though open data is not novel, individuals unfamiliar with it and its ramifications may find it intimidating. Also, the present level of interest in unprocessed open data might need to be clarified. As a result, this open data's value and potential applications are underscored by (Attard et al., 2015). The process of preparing and publishing open data poses particular challenges. Participants complained about weak government interaction with the OGD initiative. They explained that this weakness stems from a need for more awareness among government entities of its importance. To address these challenges, recommendations have been made to the entity responsible for developing a national data framework that encompasses all the support needed to manage data issues between different parties (Rashideh et al., 2022).
	Motivation	The offered data may be regarded as superfluous with no apparent purpose. An effective motivator for this endeavor is to assist data providers in comprehending the significance of the data they generate by showcasing its unforeseen applications (Zine et al., 2022).
	Capacity	Open data should be accessible to all individuals or automated systems for utilization, adaptation, and dissemination. Regrettably, data providers prioritize disseminating data only after adequately guaranteeing its high quality. Consequently, the released data has no value. Urgent implementation of standards and extensive training on a large scale are necessary for this (Zine et al., 2022).
	Budget provision	To stimulate the OGD initiative, it is imperative to allocate a specific budget and prioritize OGD initiatives to achieve their full potential.
	Technical Support	Most of the existing OGD portals were not planned to publish and consume open data on a large scale. For this reason, public entities need technical support to update their portals so that the data they provide can reach maximum reuse potential (Zine et al., 2022).
	Institutionalization	Establishing an institutional structure for the OGD initiatives would aid in determining the required responsibilities.

The majority of existing works do not address the challenges that prevent entities from joining this initiative or take into account the perspectives of data providers regarding protection and security, such as (Zine et al., 2022), (Valli Buttow & Weerts, 2022), (Attard et al., 2015).

In the next section, we provide guidelines for data providers considering data privacy protection and security.

4. Data Protection Guidelines for Publishing OGD

In this section, we focus on the third-dimension category, which corresponds to the challenges that discourage entities from joining the OGD initiative since this category has yet to be efficiently considered in current literature with severe guidelines.

Data providers have many challenges that hinder their participation in the Open Data Governance (OGD) effort. These include the spreading of information that goes against specific laws, the violation of protecting trade secrets, the invasion of privacy, the danger to infrastructure security caused by exploiting complex infrastructure data, and the revealing of inappropriate data or information. These issues could lead to negative publicity or condemnation from other public sector organizations, highlighting the need for comprehensive measures and supervision to protect the data.

We offer technical guidelines to data providers on protecting and securing their data privacy. These guidelines are based on an analysis of existing literature (Choenni et al., 2022); (Rashideh et al., 2022). We have gathered potential solutions and organized them systematically to create practical strategic protection guidelines.

It is important to note that laws regarding data can vary widely across jurisdictions, and what may be considered illegal in one location might not be in another. Organizations and individuals are responsible for understanding and complying with the relevant legal frameworks governing data in their specific context. Violations of data-related laws can lead to legal consequences, including fines, penalties, and legal actions.

1. Assessing data for compliance with legislation before publication and identifying potential legal barriers;

2. When assessing the situation, it is essential to consider the possible privacy threats. It is also possible that there may be other sorts of sensitive information that need protection, such as trade secrets or classified information;

3. Apply controls to datasets before publication to evaluate whether a dataset meets all defined privacy criteria. These controls could help reduce potential risks;

4. If specific data publication is forbidden for privacy or other reasons, publishing anonymized data may be an option;

5. Providing a data catalog containing metadata describing the datasets to be published for a better interpretation;

6. Establishing the terms and conditions for utilizing Open Government Data (OGD) datasets by limiting their usage limitations. However, the terms of use may also contain disclaimers that allow the OGD publisher to inform potential users that it is not responsible for any mistakes or omissions in the data.

5. Conclusion

The current literature addresses and discusses the different types of issues and challenges that the OGD initiative faces in reaching its full potential. However, guidelines have been provided to address OGD publishing issues to enable stakeholders to exploit published OGD and motivate developers to create innovative applications. This research analyzed the guidelines for OGD publication and classified the challenges preventing this initiative from reaching its full potential to focus on the data protection strategic dimension. Finally, we proposed a set of technical guidelines considering the issues extracted by data providers when safely joining this initiative. The objective was to minimize data protection and security disparity and incentivize data suppliers to share their data safely.

As part of future work, we aim to explore the possibilities of AI (Artificial Intelligence) within the framework of OGD platforms by conducting experimental studies and measuring its effects on the OGD initiative.

References

[1] Attard, Judie, Fabrizio Orlandi, Simon Scerri, and Sören Auer. "A systematic review of open government data initiatives." Government information quarterly 32, no. 4 (2015): 399-418.

[2] Berners-Lee, Tim. "Is your linked open data 5 star." Berners-Lee, T. Linked Data. Cambridge: W3C (2010).

[3] Bouchelouche, Khadidja, Abdessamed Réda Ghomari, and Leila Zemmouchi-Ghomari. "Open Government Data (OGD) Publication as Linked Open Data (LOD): A Survey." International Journal of Computer and Information Technology (2279-0764) 10, no. 1 (2021).

[4] Bouchelouche, K., Ghomari, A. R., & Zemmouchi-Ghomari, L. (2022). Enhanced analysis of Open Government Data: Proposed metrics for improving data quality assessment. 2022 5th International Symposium on Informatics and its Applications (ISIA), pp. 1-6.

[5] Choenni, S., Bargh, M. S., Busker, T., & Netten, N. (2022). Data governance in smart cities: Challenges and solution directions. Journal of Smart Cities and Society, 1 (1), 31-51.

[6] Gottfried, A., Hartmann, C., & Yates, D. (2021). Mining open government data for business intelligence using data visualization: A two-industry case study. Journal of Theoretical and Applied Electronic Commerce Research, 16 (4), 1042-1065.

[7] Liu, Q., Bai, Q., Ding, L., Pho, H., Chen, Y., Kloppers, C., et al. (2011). Linking Australian government data for sustainability science case study. In Springer (Ed.), Linking Government Data (pp. 181–204).

[8] López Reyes, M. E., & Magnussen, R. (2022). The Use of Open Government Data to Create Social Value. In International Conference on Electronic Government (pp. 244-257).

[9] Quarati, A. (2023). Open government data: usage trends and metadata quality. Journal of Information Science, 49 (4), 887–910.

[10] Rashideh, Waleed, Maram Alajmi, Abdulaziz Alshammari, Anas Alqudah, Waeal J. Obidallah, and Mohammed Alkhathami. "Investigating the challenges and value creation of open government data initiatives." IJCSNS 585 (2022).

[11] Šlibar, B., Mu, E.(2022). OGD metadata country portal publishing guidelines compliance: A multi-case study search for completeness and consistency. Government Information Quarterly, 39 (4), 101756.

[12] Solar, M., Concha, G., & Meijueiro, L. (2012). A model to assess open government data in public agencies. In International Conference on Electronic Government (pp. 210-221). Springer.

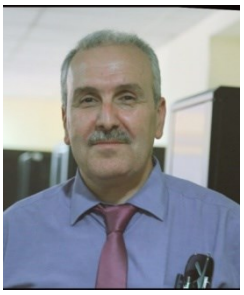
[13] Spalević, Ž., Milic, P., Veljković, N., Milićević, V., Milosavljević, S.(2023). Utilization of Open Government Data for Environmental Protection: OGD IN ENVIRONMENTAL PROTECTION. Journal of Scientific & Industrial Research (JSIR), 82 (07), 711-720.

[14] Valli Buttow, C., & Weerts, S. (2022). Open Government Data: The OECD's Swiss army knife in government transformation. Policy & Internet, 14 (1), 219-234.

[15] Zine, Hocine, Kheir Eddine Medkour, Leila Zemmouchi-Ghomari, and Abdessamed Réda Ghomari. "Open Data Influence on Digital Governance." International Journal of Innovation in the Digital Economy (IJIDE) 13, no. 1 (2022): 1-11.

Authors' biographies

KhadidjaBouchelouche a Ph.D. student in Computer Science from ESI (Ecole Nationale Supérieure d'Informatique), Algiers. She is affiliated with the LMCS (the Laboratory of Systems Design Methodologies). Her research interests include Linked Open Data, Semantic Web, and Ontology Engineering.



AbdessamadRédaGhomari received his Ph.D. in Computer Science from ESI (Ecole Nationale Supérieure d'Informatique), Algiers, in 2008. He is currently a Full Professor at ESI, Algiers, Algeria. Since 2001, he has been the Information System Management Team Head at LMCS (the Laboratory of Systems Design Methodologies). His research interests focus on knowledge engineering, Crisis Informatics, Open Government Data, and Data Governance. He has supervised the activities of partnerships with enterprises and continuing training (2011-2018).

Leila Zemmouchi-Ghomari received her Ph.D. in Computer Science from ESI (Ecole Nationale Supérieure en Informatique), Algiers, in January 2014. Her research interests include Ontology Engineering, Knowledge Engineering, Semantic Web, Linked Open Data, and Industry 4.0. She is an associate professor at the Industrial Engineering and Maintenance Department at ENSTA: Ecole Nationale Supérieure des Technologies Avancées, Algiers, Algeria.