Open Government Data (OGD) Publication as Linked Open Data (LOD): A Survey

Khadidja Bouchelouche
LMCS, Ecole nationale Supérieure d'Informatique, ESI
Algeria
Email : k_bouchelouche [AT] esi.dz

Abdessamed Réda Ghomari
LMCS, Ecole Nationale Supérieure d'Informatique, ESI
Algeria
Email : a_ghomari [AT] esi.dz

Leila Zemmouchi-Ghomari
Ecole Nationale Supérieure de Technologie, ENST
Algeria
Email : leila.ghomari [AT] enst.dz

Abstract—**Open Government Data (OGD) is a movement that has spread worldwide, enabling the publication of thousands of datasets on the Web, aiming to concretize transparency and citizen participatory governance. This initiative can create value by linking data describing the same phenomenon from different perspectives using the traditional Web and semantic web technologies. A framework of these technologies is linked data movement that guides the publication of data and their interconnection in a machine-readable means enabling automatic interpretation and exploitation. Nevertheless, Open Government Data publication as Linked Open Data (LOD) is not a trivial task due to several obstacles, such as data heterogeneity issues. Many works dealing with this transformation process have been published that need to be investigated thoroughly to deduce the general trends and the issues related to this field. The current work proposes a classification of existing methods dealing with OGD-LOD transformation and a synthesis study to highlight their main trends and challenges.**

Keywords: Open Government Data, Linked Open Data, Transformation approaches

I. INTRODUCTION

Open Government Data (OGD) has spread all over the world among many web platforms. OGD is defined as data generated by the publicly available government and can be freely shared, modified, and used for any purpose [1]. OGDs have many benefits, such as improving transparency and accountability, improving public services' quality and efficiency, promoting citizen participation, and increasing economic opportunities. However, OGD applications often require the integration of data from different data sets.

One way to solve the heterogeneity problem of these different datasets is the Linked Open Data (LOD) movement. LOD is a growing movement that aims to publish data in the most interoperable and most productive way so that individuals and institutions can use it to create exciting applications and perform in-depth analyses.

The online publication of thousands of datasets creates added value by linking several data describing the same phenomenon from different angles. Combining large-scale data is possible using current technologies, with standards like URI, RDF, and OWL. Thus, the principles of linked data favor data publication and their interconnection in a machine-readable way using Web standards.

Linked data, along with other technologies in the Semantic Web, allows data to be interconnected and reused across organizational boundaries instead of being data silos used by a single organization. For example, Facebook introduced its Open Graph Protocol in 2010 to manage user preferences using vocabularies from the cloud of linked data such as SKOS and FOAF. If someone likes a Facebook page, those friends will probably like it too. It should also be noted that more than 100,000 websites use this protocol. Also, the BBC Nature site collects data from several sources (e.g., the IUCN Red List of Threatened Species, WWF WildFinder, Wikipedia, the Animal Diversity Web, and the EDGE of Existence program of the Zoological Society of London) and reuse them in a BBC context [2].

However, this is not a trivial task because of the issues of data heterogeneity. As an example, geospatial data in multiple data sets can be represented in a different format using an administrative division or a geographic coordinate system (longitude and latitude), which can also be at a level of granularity different from the highest division to the lowest division (regions, provinces, districts).

Many works dealing with OGD-LOD transformation have been published that need to be investigated thoroughly to deduce the general trends, issues related to this field. This work aims to reuse, readjust, and exploit the data of OGD in an inventive and efficient way that tackles some of the issues of the existing literature. The current work follows a systematic literature approach that covers and considers existing works of transforming OGD to LOD with a rigorous selection.

This paper's organization is as follows: section II presents the basic concepts of the paper domain and describes the followed methodology in this work. Section III presents the study of existing approaches of OGD_LOD transformation by identifying their main contributions and limitations. Section IV describes the proposed approaches classification based on the joint contributions and the discussion resulting from this analysis. Section V is dedicated to highlighting the findings of this research and its perspectives.

II. BACKGROUND

In this section, we introduce the basic concepts of the paper domain: open government data and open linked data and the followed methodology.

A. Open Government Data

Open data is the data that is freely accessible to the public and must not discriminate against anyone [13]. Therefore, the data released in a format of open data is supposed to be "platform-independent, machine-readable and made available to the public without restrictions that would hinder the reuse of that information."

OGD is a subset of open data. Quite simply, it is government data that is freely available to the public [14]. Several datasets could belong to government data, comprising data owned by public administrations indirectly (example via agencies or subsidiaries), as data associated with climate/pollution, child care/education, congestion/traffic, public transportation [13].

B. Linked Open Data

The term linked data is defined as a set of best practices for the publication of structured data on the Web. Tim Berners-Lee has coined these principles. The principles are [15]:

- Use URIs as names for things,
- Use HTTP URIs so that people can look up those names,
- When someone looks up a URI, provide useful information (RDF/SPARQL),
- Include links to other URIs so that they can discover more things.

Linked data, along with other technologies in the Semantic Web, allows data to be interconnected and reused across organizational boundaries instead of being data silos used by a single organization. As an example of a linked data application, Facebook introduced its Open Graph Protocol in 2010 to manage user preferences using vocabularies from the cloud of linked data such as FOAF and SKOS. For example, if someone likes a Facebook page, those friends will probably like it too. It should also be noted that more than 100,000 websites use this protocol. The BBC Nature site also collects data from different sources and reuses them in a BBC context [2].

C. Methodology

In order to be able to cover most prominent the existing works of the transformation of OGD to LOD, analyze and study them in an organized and detailed manner, we followed a systematic literature approach which allows us to meet our needs explicitly and systematically, based on the guidelines proposed in [3] and [4] with explicit inclusion and exclusion criteria. The detailed steps are presented as follows:

a) Defining the research questions

- What are the existing approaches of publication and exploitation of LOD from OGD, and how can they be classified?
- What are the existing approaches allowing the transformation of OGD to LOD, and how can they be classified?
- Are there any defined guidelines for the transformation of OGD to LOD?
- What are the existing challenges related to the transformation of OGD to LOD?

b) Search strategy

To cover the broadest possible range of relevant publications, we have used the most extensively used electronic libraries and search engines, namely: Elsevier/ Springer/ IEEE Xplore /Science direct/ HAL, Semantic scholar engine/ Google scholar engine.

A list of search terms of this field was used, which are:

- Transformation of open government data to linked open data.
- Transformation of OGD to LOD.
- From open government data to linked open data.
- From OGD to LOD.

Even though the existence of the search terms in the title or the abstract of the publications but does not guarantee that all the results are relevant for our research questions, this is why the selection of the relevant results was made by applying inclusion and exclusion criteria as following:

Inclusion criteria: Publications that meet one of the following inclusion criteria are selected as primary studies:

- A study that focuses on linked open government data.
- A study that focuses on the transformation of OGD to LOD.
- A study that focuses on publishing or consuming government data as LOD.

Exclusion criteria: Publications that meet any of the following criteria are excluded from the review:

- A study that does not focus on publishing or consuming government data as linked data.
- A study that does not provide any approach/framework to facilitate the transformation of OGD to LOD.

The selection of the relevant results was made by following these steps (figure 1):

- First selection: based on scanning the title.
- Second selection: based on scanning the abstract.
- Third selection: based on scanning the whole publication.

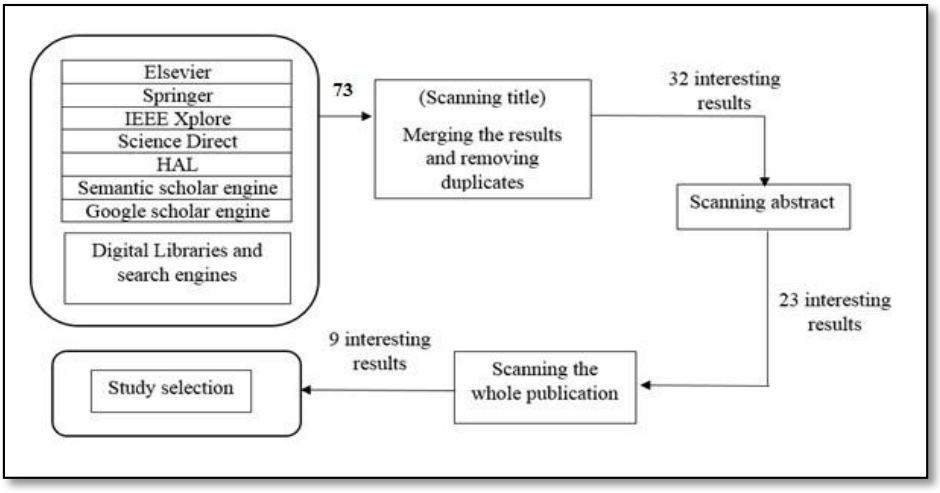

Fig. 1. The different steps of the study selection.

The next section presents the study of the selected approaches of OGD_LOD transformation by identifying the main contributions and limitations

III. LITERATURE REVIEW

We present the nine approaches of OGD-LOD transformation obtained from the previous section.

- **De Faria Cordeiro et al. approach [5]:** It aims to support the sharing, exposure, and association of data resources in the form of LOD by using a workflow management system (in the form of an Extraction-Transformation-Loading "ETL" tool) to integrate a set of tools used in the publication process and also by including a data semantics enrichment process.

  This approach offers a user-friendly environment to foster data publication and their association with other existing data. Thus, agencies publishing governmental data using this approach can easily control and monitor the publication process and increase the reuse of data transformation and publication tasks. Data consumer applications can use the integrated and standardized datasets with richer semantics based on the citizen's opinion.

  This approach faces two main limitations: 1) The lack of flexibility to add new links which stimulate associativity between resources, which can significantly contribute to increasing the number of applications compared to government data, especially if we consider new opportunities for teams of development and the private sector. 2) The lack of collaboration of citizens in the mapping and links, which is of great importance in managing provenances and the mechanisms to help the data retention process, becomes crucial and constitutes future work.

- **Boonlamp approach [6]:** This approach is based on the technology of linked data to convert relational databases to RDF, mapping the RDF data model with semantic vocabulary and ontology and convert expressions in Thailand language into SPARQL. This approach allows collaboration amongst Thailand's local governments based on its Sub-District Administrative Organization.

  The approach enables data relations that were not previously linked to provide the possibility of exposing, sharing, and linking pieces of data, information, and knowledge on the semantic Web using RDF, URLs, ontology, and data model. Also, it allows overcoming the limitation of discovery in the Thailand Local Government system and allowing to facilitate collaboration amongst Thailand's local governments and can help organizations in collaborating in strategizing and project planning.

  However, this approach is limited to the case of Thailand's local governments. Also, the non-application of the existing standard ontology with the data model in the mapping file.

- **Kalampokis approach [7]:** This approach is a two-phased approach that allows supporting participatory decision-making based on social and government open data analysis. The architecture relies on using a linked data paradigm as a layer to enable data integration from different sources.

As contributions, this approach enables citizens to participate in democratic decision-making and express their opinion on their favorite social media platforms without being aware that their opinions could impact the decision-making process. It also supports participatory decision-making and enables decision-makers to understand public opinion and predict their reactions to a decision. This ability will allow decision-makers to timely implement corrective actions.

As contributions of the latter, it aims to discover decentralized Semantic Web data, and an architectural element is provided which (a) creates the missing links, and (b) allows the clients of Semantic Web to find and integrate independent sub-graphs of different web sites, into a single, large graph.

However, the approach has some limitations: 1) the need to enrich the DBpedia ontology with appropriate classes and relations. DBpedia allows the user to enrich it with concepts and classes. Thus, an in-depth examination of the candidate websites of all affected countries is needed to analyze their content. As a sequel, a concept taxonomy of relations will be created to merge all semantic information contained in those websites. This taxonomy can be provided as a complement to DBpedia, which can also be utilized as a starting point for other purposes. 2) All the work involves the manual addition of the links (such as RDFS:seeAlso, OWL:sameAs). The Silk framework that can provide automatic discovery and proposing such links need to be tested.

- **Aryan et al. approach [9]:** It aims to foster public participation and government transparency by linking open government data by converting existing data to RDF format and access to the integrated data.

  As contributions, this approach served as a lens to magnify the range of interaction possibilities with the data to foster public participation, as it allows to stimulate public participation and government transparency.

  For the limitations: 1) The current prototype is only dedicated to the case study's specific data. 2) It is necessary to minimize the storage requirements by applying the conversion on the fly. This would help the government and nurture the public participation culture in Indonesia for the open data movement.

- **Budsapawanich et al. approach [1]:** It represents a conceptual framework to map and link OGD datasets using geospatial data. The framework generates geospatially-linked OGD from structured and semi-structured datasets.

  As contributions, the framework provides automatic generations of links among OGD datasets using geospatial data with no need for intensive human workloads.

  For the limitations, 1) the framework is not suitable for non-geospatial fields. 2) The framework needs to be expanded to support other types of potential linking nodes, such as time. 3) The framework does not provide linked data visualization and insightful data analysis.

- **Krataithong et al. approach [10]:** this approach represents a framework for supporting automatic linking RDF datasets by looking for semantic types in CSV datasets. The semantic type information will be used to map the related name entities to URIs in the resulted RDF dataset. The two crucial steps that are focused on in this paper are the process for semantic type detection and URIs lookup for name entities.

  As contributions, the framework supports the RDF datasets' automatic linking for some datasets of the Thailand portal. The proposed approach can also enhance coverage while mapping Thailand name entities and can resolve the word ambiguity problem, which are two main problems in identifying the proper URIs for name entities.

  As limitations, the proposed framework is limited to the case of the Thailand open government data portal.

- **Trinh et al. approach [11]:** It aims to eliminate the barrier between end-users and Open Government Linked Data to exploit linked data effectively. The underlying idea is the Linked Widgets Platform (LWP), which is composed of several connectable and straightforward applications. Each application fulfills a simple task and can be linked to enabling new, more complex tasks.

  As contributions, the architecture is flexible and dynamic, with an open system that can be extended as desired. Also, it allows all users to exploit linked data using simple operations, and it gives the possibility of creating new applications. As a result, the solution providers and novice users can create and adapt individual linked data applications with a user-friendly environment with no worries about the technical challenges of consuming linked data.

  The existing limitation is one of the platforms that need to be implemented with additional useful functionality to allow users only to describe what they desire, and the

  As limitations, this approach lacks a conceptual model that allows describing the join points of social and government data realities and the need for implementing the proposed architecture and identify real-world use case scenarios to evaluate both the proposed approach and the foreseen system.

- **Galiotou and Fragkou approach [8]:** This approach represents a methodology for discovering decentralized Semantic Web data.

system will provide the widgets that can be used or can automatically create a mashup for them.

- **Vafopoulos et al. approach [12]:** It represents a conceptual model named "Linked Open Economy (LOE)" that allows the exploitation of enormous amounts and diversity of open economic data that are progressively becoming available by open source communities and governments. The primary purpose is to release the power of open source systems and build a common ground to catalyze, providing more realistic answers in significant economic activities.

  As contributions, LOE can be used to (i) allows exchanging information between open source systems, (ii) integrate open data from heterogeneous sources, and (iii) publish semantically and linked data about economic activities. It can also be connected to market processes and provide tremendous valuable insights to other domain stakeholders (e.g., developers, citizens, journalists, researchers, companies) who can use it as a compact common ground to have and compile valuable economic information.

  As a limitation, the approach is limited to the case of economic data.

After a deep understanding of these nine approaches, we deduce that the latter can be classified that meet close contributions and purposes.

IV. PROPOSED APPROACHES CLASSIFICATION

In this section, we present the proposed classification of the existing approaches in the field of OGD-LOD Transformation and discuss the results of this study.

A. Classification

After studying and analyzing the existing approaches in the field of the OGD-LOD transformation (previous view section), we noticed that the existing works could be classified into four categories according to the provided contributions and based on the following criteria:

- Supported approaches of publishing OGD as LOD,
- Level of end-users' interaction with government data.
- Degree of linked OGD efficiency based on combining multiple government data to form a unified network of government-linked data,
- Effectiveness of linking the RDF dataset to form linked data.

TABLE I. THE CLASSIFICATION OF THE CURRENT APPROACHES INTO CATEGORIES

Category	Publishing OGD as LOD	Transforming OGD into LOD	Exploiting OGD as LOD	Linking RDF datasets
Description	This category includes the studies that contain approaches enabling the publishing of OGD as LOD to support the sharing, exposure, and association of data resources as LOD.	This category includes transforming multiple published government data (resulted from the first category) to form a unified network of linked data, which will enable linking the maximum of data of the same city, country, or countries.	This category includes approaches that aim to improve the level of interaction of end-users with government data by allowing the end-users to specify their needs to use government data to obtain the appropriate network of linked OGD. This category is dynamic, and its output can be modified according to the end-users' needs. This is why it is for the specified case of exploiting OGD as LOD.	This category's approaches allow linking RDF datasets effectively to form linked data, since linking RDF datasets is considered a fundamental condition to reach LOD from OGD. Hence, this category can be considered a part of the three previously mentioned categories that promote the process of linking RDF datasets.
Issues	It does not guarantee that the published OGD is adequate for end-users' needs.	The difficulty arises while combining different data types because each OGD portal has its data types, such as TXT, CSV, PDF, XML, and JSON.	The lack of a concrete application allows users to exploit OGD as they are supposed to and according to their needs.	The lack of a standard toolkit that allows linking RDF datasets automatically.
Works	[5], [6]	[1], [7], [8], [9]	[11], [12]	[10]

B. Discussion

This study allows us to investigate the different ways of obtaining government data in the form of linked data, which is very interesting for stakeholders to exploit OGD data in practical ways and enable OGD initiatives to reach their full potential.

The findings methods that allow OGD-LOD transformation can be categorized into four classes based on the following criteria, (1) the supported approaches of publishing OGD as LOD, (2) the level of end-users' interaction with government data, (3) the degree of linked OGD efficiency based on combining multiple government datasets to form a unified network of government-linked data and (4) the effectiveness of linking RDF dataset to form linked data.

This classification allows us to bring together the approaches that meet in contributions that will enable future work to focus on a specified category and to propose appropriate solutions for them.

Conventional boundaries that can be seen as challenges are:

- the need for a standard and generic approach that is dedicated for any case study and any data domain (for example, economic, health, education),
- the need for a platform that allows users to input and search data,
- the lack of citizens' collaboration in the mapping and identifying links between datasets,
- the need for methods and tools that allow the automatic linking/integration of data without requiring intensive human workload,
- the need for tools to visualize linked data and provide insightful data analysis.

Proposed solutions to these challenges converge to providing a generic approach to transforming data to RDF triples and automatically linking them. Then, providing a sophisticated platform for end-users that enables citizens' collaboration to update and correct data. Besides, providing visualization tools for data exploitation and analysis.

V. Conclusion

We have followed a systematic literature review to study and analyze the several approaches to transforming OGD into LOD.

This analysis led us to classify the selected approaches that meet close contributions and purposes, as mentioned in the discussion.

This research allows us to answer the research questions mentioned in section 2, so we discovered the current publication and exploitation of LOD from OGD and classified them. We also identified the linked data principles and technologies as promising guidelines for publishing OGD as LOD. Finally, we studied the challenges related to the transformation of OGD into LOD.

This research allowed us to discover and understand the tendency of the current OGD-LOD transformation works, which are specific and, therefore, difficult to reuse.

As future work, we plan to propose a generic approach that aims to automatically link/integrate RDF datasets in order to tackle the issue of requiring intensive human workloads as well as supports the collaboration of citizens in the mapping and links, which is of great importance in establishing useful links in the purpose of facilitating the use of OGD in inventive ways.

This perspective will be concretized within the research project entitled OGDIVAA (Open Government Data Initiatives and Value delivery for Algerian public Agencies) approved under the number C00L07ES160520200004 by the Algerian Ministry of Higher Education and Scientific Research.

References

[1] P. Budsapawanich, C. Anutariya, and C. Haruechaiyasak, "A Conceptual Framework for Linking Open Government Data Based-On Geolocation: A Case of Thailand," Springer, 2018, pp. 352-366 [Joint International Semantic Technology Conference].

[2] L. Zemmouchi-Ghomari, "Linked Data: A Manner to Realize the Web of Data," Chapter V, in Handbook of Research on Technology Integration in the Global World, 2019, pp. 87-113.

[3] B. Kitchenham, "Procedures for performing systematic reviews," Keele, UK, Keele University, vol. 33, pp. 1-26, 2004.

[4] T. Dyba, T. Dingsoyr, and G. K. Hanssen, "Applying systematic reviews to diverse study types: An experience report," IEEE, 2007, pp. 225-234 [First International Symposium on Empirical Software Engineering and Measurement (ESEM 2007)].

[5] K. De Faria Cordeiro, F. F. de Faria, B. de Oliveira Pereira, A. Freitas,E. Ribeiro, J. V. V. B. Freitas, A. C. Bringuente, L. de Oliveira Arantes, R. Calhau, and V. Zamborlini, "An approach for managing and semantically enriching the publication of Linked Open Governmental Data," 2011, pp. 82-95 [Proceedings of the 3rd workshop in applied computing for electronic government (WCGE), SBBD].

[6] L. Boonlamp, "A linked data approach to planning collaboration amongst local governments in Thailand," IEEE, 2017, pp. 1-5 [chez 2017 2nd International Conference on Information Technology (INCIT)].

[7] E. Kalampokis, M. Hausenblas et K. Tarabanis, "Combining social and open government data for participatory decision-making," Springer, 2011, pp. 36-47 [International Conference on Electronic Participation].

[8] E. Galiotou et P. Fragkou, "Applying linked data technologies to Greek open government data: a case study," Procedia-social and behavioral sciences, vol. 73, pp. 479-486, 2013.

[9] P. R. Aryan, F. J. Ekaputra, W. D. Sunindyo, and S. Akbar, "Fostering government transparency and public participation through linked open government data: Case study: Indonesian public information service," IEEE, 2014, pp. 1-6 [2014 International Conference on Data and Software Engineering (ICODSE)].

[10] P. Krataithong, M. Buranarach, N. Hongwarittorrn, and T. Supnithi, "A framework for linking RDF datasets for Thailand open government data based on semantic type detection," Springer, 2016, pp. 257-268 [chez International Conference on Asian Digital Libraries].

[11] T.-D. Trinh, B.-L. Do, P. Wetz, A. Anjomshoaa, and A. M. Tjoa, "Linked widgets: An approach to exploit open government data," 2013, pp. 438-442 [Proceedings of International Conference on Information Integration and Web-based Applications & Services].

[12] M. Vafopoulos, S. Rallis, I. Anagnostopoulos, V. Peristeras, D. Negkas, I. Skaros, and A. Tzani, "Mining and Linking Open Economic Data from Governmental Communities," Springer, 2018, pp. 144-148 [IFIP International Conference on Open Source Systems].

[13] J. Attard, F. Orlandi, S. Scerri, and S. Auer, "A systematic review of open government data initiatives," Government Information Quarterly, 2015.

[14] J. Kučera, D. Chlapek, and M. Nečaský, "Open government data catalogs: Current approaches and quality perspective," Springer, 2013, pp. 152-166 [International conference on electronic government and the information systems perspective].

[15] C. Bizer, T. Heath et T. Berners-Lee, "Linked data — The story so far," *International Journal Semantic Web Information Systems,* vol. 5, n° %13, pp. 1-22, 2009.