Hindawi
Journal of Electrical and Computer Engineering
Volume 2023, Article ID 4854965, 17 pages
https://doi.org/10.1155/2023/4854965

Research Article

The Relevance of Open Data Principles for the Web of Data

Jhon Francined Herrera-Cubides,[1] Paulo Alonso Gaona-García,[1] Carlos Enrique Montenegro-Marin ,[1] and Salvador Sánchez-Alonso[2]

[1] *Universidad Distrital Francisco José de Caldas, Bogotá, Colombia*
[2] *Universidad de Alcalá, Alcalá de Henares, Madrid, Spain*

Correspondence should be addressed to Carlos Enrique Montenegro-Marin; cemontenegrom@udistrital.edu.co

Received 21 March 2023; Revised 22 July 2023; Accepted 2 August 2023; Published 14 September 2023

Academic Editor: Nihal F. F. Areed

Open data has been improving both publishing platforms and the consumers-oriented process over the years, providing better openness policies and transparency. Although organizations have tried to open their data, the enrichment of their resources through the Web of Data has been decreasing. Linked data has been suffering from notable difficulties in different stages of its life cycle, becoming over the years less attractive to users. According to that, we decided to explore how the lack of some opening requirements affects the decline of the Web of Data. This paper presents the Web of Data radiography, analyzing the governmental domain as a case study. The results indicate that it is necessary to strengthen the data opening process to improve resource enrichment on the Web and have better datasets. These improvements describe that open data must be public, accessible (in machine-readable formats), described (use of robust, granular metadata), reusable (made available under an open license), complete (published in primary forms), and timely (preserve the value of the data). The implementation of these characteristics would enhance the availability and reuse of datasets. Besides, organizations must understand that opening and enriching their data require a completely new approach, and they have to pay special attention and control to this project, generally by putting money, the commitment by management at all levels, and lots of time. On the contrary, given the magnitude of availability and reuse problems identified in the opening and enrichment data process, it is believed that the Web of Data model would inevitably lose the interest it aroused at the beginning if not addressed immediately by data quality, openness, and enrichment issues. Besides, its use would be restricted to a few particular niches or would even disappear altogether.

1. Introduction

Linked open data (LOD) is an initiative suggested by organizations to make their data available in a machine-readable format. This requirement allows users to use and combine available datasets to create knowledge and apps in their context [1]. In addition, LOD has been working with two major concepts: linked data and open data. On the one hand, linked data defines a set of design principles for adding value to data by linking to other data (data enrichment). On the other hand, data available under a given license for use, reuse, and redistribution by any person or organization [2] are called open data. To carry out this proposal, the authors in [3] proposed a linked data 5-star scheme. Although the 5-star scheme has a lot of advantages, two relevant problems have been recognized by the authors in [4, 5]. Firstly, most open data systems do not manage dataset reuse completely, even though datasets are available (1–3 linked data levels). This lack of reuse (replicating and redundancy of existing data) does not allow for interlinking among existing data, decreasing the possibility of creating a richly interconnected data network on the Web of Data. Secondly, not all linked data is open data, and not all open data can be linked. In the Web of Data, the licensing terms of the published datasets determine whether the data are freely available and open for anyone to use, reuse, share, and distribute.

Researchers such as [6, 7] have identified the rapid growth in the quantity of LOD repositories, which use platforms such as comprehensive knowledge archive network (CKAN) [8–10] to manage their services. These platforms let publish and exploit datasets and metadata. However, despite data opening and linking guidelines [1, 11–13], there are challenges in different linked data lifecycle stages. These challenges can affect the opening and enrichment of data available on the Web of Data. Some issues described by the authors of [14–16] are (1) the lack of use of machine-readable data formats, appropriate data license terms, provenance and quality attributes, data vocabularies, and data access strategies [17]. These problems hamper the freely available and open data; (2) the lack of apps to detect possible data quality issues [18–21], such as inconsistency, inaccuracy, out-of-date, and incompleteness; (3) the reliability of the search results is defined by the reliability of the datasets from which these results were obtained [22]. For example, if you link datasets that have inconsistency problems, you would not lend value added to your data. Finally, (4) data on the Web show a significant data quality variation. For example, data extracted from semistructured sources, such as DBpedia, often contain inconsistencies and false and incomplete information.

According to this context, we decided to explore the current status of the Web of Data, focusing on the main opening requirements defined by the Linked Data guidelines. Studies such as [4, 17, 23] and [24] propose the criteria and methods used in this research. For this aim, this analysis works on two main approaches: firstly, the requirements to reach the dataset availability, and secondly, the necessary information to achieve the dataset reuse. These approaches allowed for the assessment of the behavior of the opening and linking processes provided by datasets published on the Web. The contribution of this paper is analyzing the main challenges of open and linked data in the Web of Data. This analysis will allow us to identify how to improve the openness and linking of our data published on the Web and finally add value to them. For this aim, this research proposes the following study questions: What are the most common issues that arise from exploit datasets published under LOD principles? How do these findings affect the decline of the Web of Data? What are the challenges addressed in linking resources under LOD principles? To solve these questions, Section 2 examines the background of the problem. Section 3 presents the methodological design and its corresponding implementation. Sections 4 and 5 review, analyze, and discuss the evidence found on the status of the Web of Data. And finally, Section 6 presents conclusions and future work.

2. Background

Different problems have reduced the use of these design principles for sharing machine-readable interlinked data on the Web. LOD suffers from serious issues such as the lack of availability of data published on the Web, the lack of use of machine-readable and reusable formats, datasets are not available free of charge and do not have openly licensed, the datasets are not up-to-date, it is not easy to find information (metadata) about these datasets, and some of these datasets have inaccuracy, incompleteness, and inconsistency issues. The Open Data Barometer [25] describes some of these issues deeply. These problems do not allow us to add semantic value to our data or link and reuse them in other contexts.

The data quality is usually understood as fitness for use. Data quality may depend on several quality dimensions. Some of these dimensions are accuracy, timeliness, completeness, relevancy, objectivity, believability, and understandability, among others, cited by Zaveri et al. [16]. In addition, data quality problems can strike at the potentiality of data applications. The lack of data provenance information is a problem in data quality evaluation [16, 26], for instance. In this sense, global data management [27] identified that human errors, too many data sources, and inadequate data strategy are the most significant issues concerning the lack of data quality. In short, the data were often from multiple heterogeneous sources, and sometimes, these sources have different quality levels [28]. Thus, data quality is the main challenge in linked data.

According to the circumstances described previously, we considered it relevant to study the first stages of the linked data life cycle, owing to the components of the opening and reuse abilities (the two proposed approaches in this research) starting in these first stages.

This research will allow us to identify other features that data must meet to reach all linked data levels and, hence, a better-linked data quality level. For that reason, this study has taken as a reference the analysis of datasets published in different instances of CKAN [8, 10] to analyze the status of the Web of Data. For that purpose, we address two specific topics: (1) previous studies of the Web of Data status and (2) challenges identified on it.

2.1. Previous Work Analyzing the Status of the Web of Data. Regarding the state of the Web of Data, different studies [4, 7, 29–36] and [37] expose several issues regarding requirements to reach all linked data levels. The low use of machine-readable formats, the lack of adequate open licensing terms which do not impede its reuse for free, metadata with little human readability, out-of-date data, the extra effort required to get the five stars of the linked data model, and the under reuse and enrichment of data briefly summarize their main findings on these researches.

The findings described above show that data quality problems are a persistent challenge in linking processes. These problems can be observed both at the level of linked data principles and the attributes abstraction that describes an addressable resource.

2.2. Web of Data Challenges. It is recommendable to meet a set of best guidelines, such as those described in [38], to discover datasets and facilitate data integration from different data sources. Despite the existence of these best practices, linked data faces challenges based, for the most part, on data quality. The authors of [17, 23] show a compilation of data on the Web challenges summarized into categories. These challenges focus on metadata, data license, provenance, quality, data versioning, data identification, data format, data vocabularies, data access, data preservation, feedback, data enrichment, and data republication.

Based on this background, this research aims to carry out the radiography of the linked resources. For this purpose, this analysis works two main criteria: availability and reuse of published data on the Web. For this analysis, this research does not seek to review the social, political, economic, or contractual issues that affect the low quality of open and linked data. On the contrary, we analyze six basic technical requirements for openness and data enrichment. As a result, we present a set of recommendations to improve the steady decline suffered by the Semantic Web. To develop this analysis, the approach of this study is further detailed in the next section.

2.3. Research Approach. The LOD model establishes a five-level schema for linked data (5 stars). Each level adds features that data must meet to reach a level of linkage. The inventor of the World Wide Web and the creator and advocate of the Semantic Web and Linked Data, Sir Tim Berners-Lee, laid down the four design principles of linked data [39]:

(a) Use URIs as names for things

(b) Use HTTP URIs so that people can look up these names

(c) When someone looks up a URI, provide useful information using the standards (RDF, SPARQL)

(d) Include links to other URIs so that they can discover more things

These principles suggested a 5-star deployment scheme for open data [40]:

1. Make your stuff available on the Web (whatever format) under an open license
2. Make it available as structured data (e.g., Excel instead of an image scan of a table)
3. Use nonproprietary formats (e.g., CSV instead of Excel)
4. Use URIs to denote things so people can point at your stuff
5. Link your data to other data to provide context

As said by Abella et al. [4], this five-level schema can be classified into two relevant aspects: availability (levels 1, 2, and 3) and reuse (levels 4 and 5). This study proposes to examine these two aspects, considering the following delimitations.

2.3.1. Availability of Opening. Having in mind that linked data defines four principles described in a 5-star scheme for linked resources, our research explores a set of requirements that support the first three linked data levels. These requirements circumscribe the necessary elements for the linked resource availability. Studies such as [2, 12, 24], and the analysis of descriptive and administrative metadata [41], allowed us to identify those elements. According to this, the components covered by this study are as follows:

(i) Publishing domain: how the datasets are naming their knowledge domain

(ii) Resource licensing: what kind of Terms of Service, attribution requirements, and restrictions on dissemination are defined

(iii) Publication format: what kind of file formats is used for data

(iv) Publications updating: how often data is being updated

These variables make it possible to identify a core process (availability), on which the opening schema and, consequently, the linked resources are supported. It is necessary to note that variables such as access, performance, or cost will not be analyzed [42] as they are variables oriented to the infrastructure service that supports linkage from a technological point of view.

2.3.2. Ability to Reuse. As exposed by Abella et al. [4], levels 4 and 5 in the LOD schema allow the reuse ability. According to that, the published data must be perfectly identifiable, able to be linked, and make its information useful to other datasets. To carry out these tasks, firstly, URIs that identify the workspace entities must be provided. Also, these URIs produce links to useful data sources, both internal and external, that enrich their data. Subsequently, the data must be published in a structured way, using the data model provided by RDF (Turtle, RDFa, and SPARQL, for instance). Regarding the RDF structure, it is based on triples (subject-predicate-object), and the objects of this triple can be a URI reference, a literal, or a blank node. Making use of the RDF structure provided by datasets, where it is specified that datasets act as linkage subjects or objects, our study analyzes the information provided by the queried datasets, to identify information concerning the reuse made of these datasets published on the Web.

Briefly, the criteria described above were selected as strategic as they contribute to establishing the open data availability and provide elements for assessing the reusability of published datasets. Also, they allow the identification of shortcomings or barriers in the process of publishing linked data and, finally, serve to identify challenges to be taken up in the linked data process.

3. Methodology

A set of stages is defined by our methodology. Firstly, and according to the Research Approach, a knowledge domain was selected (Section 3.1). After that, the data consumption process was performed using a query tool that was built for this purpose (Section 3.2). Then, the analysis of the results about resource availability and reuse was performed (Section 4). Finally, the findings obtained were analyzed (Section 5). The compilation and analysis of the entire study were carried out from November 2018 to July 2019, and the final results were acquired in January 2020.

3.1. Selection of the Knowledge Domain. To define the knowledge domain, a repositories analysis was performed. In this review, problems such as proprietary approach, login use, and behavior as storage banks, the lack of data exploitation services, and proprietary data management platforms, among others, were identified. Based on these problems, repositories that use CKAN instances were selected [8, 10]. CKAN helps users from different domains (governments, companies, and organizations), in order to publish their data through a data management workflow. CKAN is the platform that handles websites such as Datahub, European Public Data Portal, or the U.S. Government's Open Data portal [43].

Although there are knowledge domains with well-defined taxonomies and data-publishing processes, they are not open access. On the contrary, and considering that openness and transparency are mandatory for the public sector, Government data repositories were selected for this research. Then, after the revision of open data catalogs, such as Open Data Inception, DataPortals.org, and the Open Data Inventory (2018 and 2019), a random sample of 40 instances of CKAN was selected and is shown in Table 1.

For the experimental design, we proposed a simple random sample. In this sampling technique, each item in the population, and every sample size, has an equal probability of being chosen in the sample. It is complex to define the dataset population size in this study owing to dataset abundance. Considering that, we selected this random sample technique for an infinite population.

To define the sample size, there are several potential ways to decide upon the size of your sample, but one of the simplest involves using a formula with your desired confidence interval and confidence level, the estimated size of the population you are working with, and the standard deviation of whatever you want to measure in your population [44]. The most common confidence interval and levels used are 0.05 and 0.95, respectively. Since you may not know the standard deviation of the population you are studying, you should choose a number high enough to account for a variety of possibilities (such as 0.5) [45].

According to the explanation above, we selected a margin of error close to 0.25% with a confidence level of 95%. Considering that each dataset has the same probability of success or failure, the result of the estimated sample is 217.778 datasets. Open Data Portal Directories, such as DataPortals.org, Open Data Inception, and Open Data Portals (TruLibraries), allow us to identify data repositories. The sample was determined using these directories by assigning sequential values to each data portal within a population, then randomly selecting those values. After that, we added all datasets of each instance until obtaining an approximate number of the sample. As a result, and according to the statistical method, a representative sample of 226.393 datasets was selected from 40 CKAN instances (Tables 1 and 2).

Finally, we selected the CKAN platform for this study because CKAN is a powerful tool for data custodians, and all its services are available for free as part of the open-source movement. In addition, hundreds of CKAN portals live, with hundreds of thousands of datasets being used. For example, there are over 800,000 datasets on the European data portal alone [46]. Some of the CKAN users include the Humanitarian Data Exchange (managed by the United Nations), data.gov.au, data.gov (US), data.gov.uk, Open Government of Canada, and the European Data Portal [47].

3.2. Data Exploitation Strategy. Two main challenges were posed to carry out the experimental phase: how to query the selected data instances and how to tabulate and visualize the queried information, taking into account the defined variables? Visual analytics for CKAN instances Tool was built for this purpose [48]. The tool provides a series of visual analytics about the current state of the datasets queried from the different datasets published in CKAN instances. This tool provides the modules described as follows:

(a) Metadata download of CKAN instances: This module uses the API provided by CKAN for the Linked Open Data consumption and storage of data for later use.

(b) Creation of REST service: This module creates a REST server that allows connection between the front end of the tool and the data of the instances.

(c) Implementation of the machine learning module to evaluate the concordance level of the metadata labels: This module generated a consumption library for unsupervised machine learning. This technique allows us to determine the concordance level of the metadata tags corresponding to each dataset of an instance.

(d) Visual Analytics module: This module implements graphic libraries to represent the metadata analysis coming from the instances of CKAN.

This tool allowed us to select the instances to be queried, use the connection services granted by the data instance, query datasets according to the variables identified, and create analysis strategies for the queried data (Figure 1).

This tool shows the particular visualizations of the selected variables and allowed storing the obtained information in JavaScript Object Notation (JSON) files [49]. Datasets were loaded and scanned to check the existence of the availability metadata, such as format, author, and date of send out, among others, and the evaluation of their behavior as linked objects or subjects. As a result, the obtained visualizations allowed us to realize the respective analysis and the construction of the judgment. Proposals such as [34, 38, 50] were considered to build the visualizations.

4. Analysis of Results

As explained in [48], the metadata describes the dataset and specifies its content. These tags allow us to relate datasets of different instances to a specific knowledge domain. For this purpose, we implemented an unsupervised machine learning module. This module determines the accuracy level of the metadata tags depending on their description. After the data load, the machine learning module consumes and

Table 1: Queried CKAN instances.

No.	Open data instance/portal	URL	No. of datasets
1	Ecuador	https://catalogo.datosabiertos.gob.ec/	108
2	Datahub.io	https://datahub.io/	11.263
3	Croatia	https://data.zagreb.hr/	44
4	The State of Rio Grande, Brazil	https://dados.rs.gov.br/	1.086
5	The State of Alagoas, Brazil	https://dados.al.gov.br/	187
6	Repository of the African Continent	https://africaopendata.org/	3.080
7	The City of Malaga	https://datosabiertos.malaga.eu/	767
8	UK Open Data	https://data.gov.uk/	11.000
9	Salzburgerland	https://data.salzburgerland.com/	5
10	Cáceres	https://opendata.caceres.es/	89
11	The University of Bristol	https://data.bris.ac.uk/data/	12.829
12	European Data portal	https://www.europeandataportal.eu/data/dataset	100.847
13	Puerto Alegre, Brazil	https://datapoa.com.br/	97
14	New South Wales Government Public Data	https://data.nsw.gov.au/data/	1.688
15	Alberta, Canada	https://open.alberta.ca/	13.394
16	Ottawa, Canada	https://data.ottawa.ca/	152
17	The Prefecture of Recife, Brazil	https://dados.recife.pe.gov.br/	61
18	Surrey, Canada	https://data.surrey.ca/	359
19	Copenhagen	https://data.kk.dk/	243
20	Montevideo	https://catalogodatos.gub.uy/	196
21	Joint Research Centre (JRC) of the European Commission	https://drdsi.jrc.ec.europa.eu/	10.210
22	Lexington	https://data.lexingtonky.gov/	98
23	Helsinki	https://hri.fi/fi/	628
24	Brazilian Open Data Portal	https://dados.gov.br/	5.219
25	Unofficial Repository of Open Public Data of the Argentine Republic	https://datar.noip.me/	128
26	Ireland	https://data.gov.ie/	1.500
27	Data Portal Project Detection of Archaeological Residues (DART)	https://dartportal.leeds.ac.uk/	25
28	Western Australia Government Open Data Catalog	https://catalogue.data.wa.gov.au/	982
29	Romania	https://data.gov.ro/	1.134
30	Fortaleza, Brazil	https://dados.fortaleza.ce.gov.br/portal/	281
31	Slovenia	https://data.gov.sk/	1.712
32	Montreal, Canada	https://donnees.ville.montreal.qc.ca/	292
33	Dutch Government Data Portal	https://data.overheid.nl/data/dataset	12.941
34	Aragon	https://opendata.aragon.es/datos	5.388
35	Chicago Metropolitan Agency for Planning Data Center	https://datahub.cmap.illinois.gov/	455
36	The International AID Transparency Initiative	https://iatiregistry.org/	5.467
37	ECAI Data Portal	https://ecaidata.org/	1.208
38	CivicData	https://www.civicdata.io/	307
39	Minnesota	https://gisdata.mn.gov/	728
40	Japan	https://www.data.go.jp/data/en/	20.195
	Total no. of datasets		226.393

Table 2: Amount of resources per queried CKAN instance.

No.	Description	No. of resources
1	Ecuador	704
2	Datahub.io	26.337
3	Croatia	167
4	Rio Grande, Brazil	15.574
5	Alagoas, Brazil	2.111
6	Africaopendata.org	10.196
7	Malaga	1.396
8	United Kingdom	4.042
9	Salzburgerland	13
10	Caceres	603
11	Bristol University	172.368
12	European Data portal	101.000
13	Puerto Alegre, Brazil	192
14	New South Wales	3.634
15	Alberta, Canada	38.522
16	Ottawa, Canada	904
17	Recife, Brazil	632
18	Surrey, Canada	1.372
19	Copenhagen	922
20	Montevideo	819
21	Joint Research Centre	24.727
22	Lexington	196
23	Helsinki	1.105
24	Brazil	27.165
25	Argentina	447
26	Ireland	3.590
27	DART	4.409
28	Western Australia	2.636
29	Romania	20.713
30	Fortaleza, Brazil	1.242
31	Slovenia	7.261
32	Montreal, Canada	1.971
33	Holland	35.634
34	Aragón	7.257
35	Chicago	1.452
36	IAID	5.419
37	ECAI	164
38	CivicData	319
39	Minnesota	3.100
40	Japan	255.132

analyzes CKAN instances to obtain stats results. We use bar charts, cake diagrams, data tables, and other visual analytics elements to show stats results. The metadata tags analyzed were as follows:

(i) Description of the organization

(ii) Author

(iii) Licenses

(iv) Formats of dataset resources

(v) Relationships as object and subject

(vi) Resources links

So, the results are graphically represented, after a statistical process, such as

(i) Percentage of empty dataset authors tags

(ii) Number of resources with a specific format

(iii) Type of licenses for each dataset

(iv) Dispersion diagrams about concordance dataset/tags

For the data analysis, the results were segmented into two sections: the resource availability and the ability to reuse.

4.1. *The Resource Availability.*

Results related to the availability of the resources queried were analyzed. To carry out this aim, the analysis of the variables associated with licensing, data format, updating date, and related domains, were worked. The results are shown below.

4.1.1. *Licensing.*

The most common licenses used for publishing data are OGL (Open Game License) (16.8%), Creative Commons Attribution (19.1%), Non-Commercial Government License for Public Sector Information (8%), and Creative Commons Zero (5.95%). Some 11% of the queried datasets have different licensing types that are too specific for their purposes or countries. Examples of this licensing type are the United Kingdom's Crown or Canada, or organizations such as IBM or MIT. Certain countries or brands use this kind of specific license according to their purposes or needs, such as:

(i) The Open Government License is used where data collections are Crown Copyright and the Creative Commons Attribution 4.0 International License is used (when available) where data collections are copyrighted by others, for instance.

(ii) The MIT license gives express permission for users to reuse code for any purpose, sometimes even if the code is part of proprietary software. As long as users include the original copy of the MIT license in their distribution, they can make changes or modifications to the code to suit their own needs. It is one of the simplest open-source license agreements. The intent was for the text to be understandable by average users and to avoid extensive litigation, which may arise from other similar Free and Open-Source Software (FOSS) licenses (https://snyk.io/learn/what-is-mit-license/).

These different licensing types subordinate to brands, products, or countries, among other aspects, can lead to difficulties using the dataset, depending on the type of permits or restrictions each country, brand, or product laid down in the license terms.

From the licensing analysis, Figure 2 shows that 26.24% of the queried datasets do not have a specific license to determine the characteristics of their use. Some causes of this issue are: Not to specify the licensing used, or not to have processed this information in the resource metadata. As proposed in [24], the lack of the "Terms of Service" description, attribution requirements, and restrictions on dissemination, among others, act as a barrier to public use of the data. Maximal openness includes clearly labeling public and available information without restrictions on use as part of the public domain.

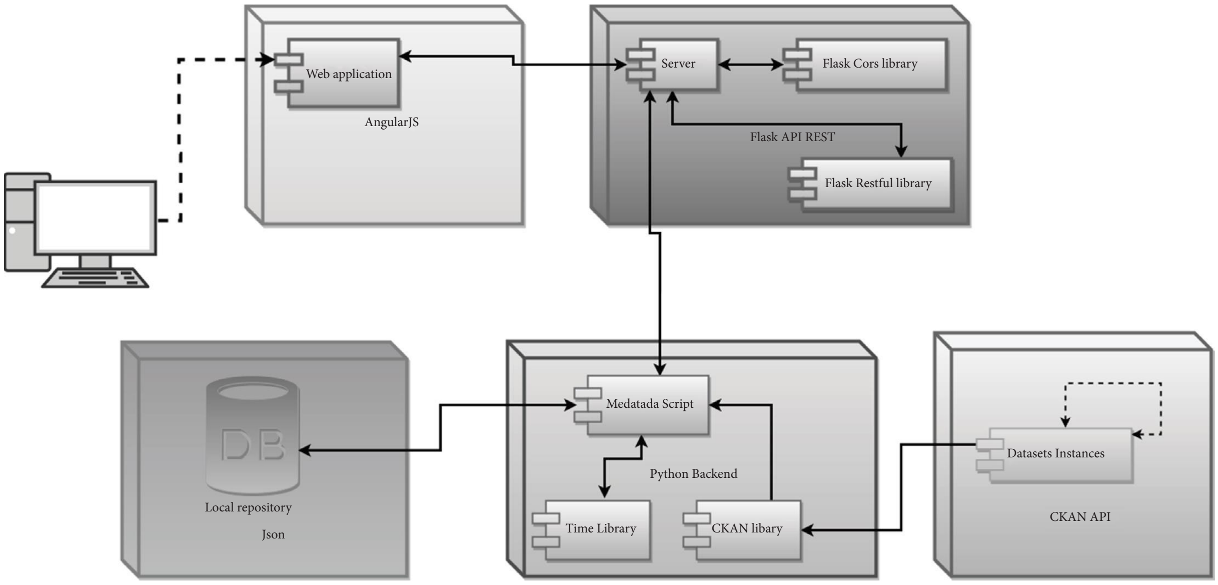

Figure 1: Technological environment of the proposed experiment.

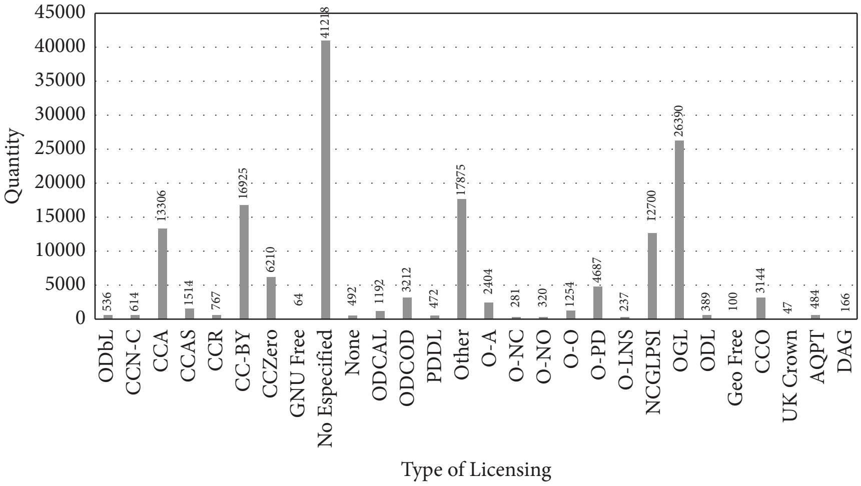

Figure 2: Most used types of licensing.

In the used generic licenses is observed a high degree of flexibility, which allows distribution, mix again and create from its work, even for commercial purposes (except for governmental ones), provided that the respective credit is given for the original creation. These types of licenses are recommended for maximum dissemination and use of licensed materials. Similarly, among other aspects, it highlights the use of attribution and noncommercial and public domain licenses, which grant the waiver of all rights to the work worldwide, and under copyright law, including all related rights to the extent allowed by law. Overall, the organizations that have entered LOD have tried to approach the 5-star scheme by publishing data on the Web (first level) but have failed to provide these resources under a clear licensing which allows actions such as reuse or redistribution.

4.1.2. Data Formats. According to the second level of the linked data scheme, CVS and XLS are the most used structured data formats. On the contrary, PDF and JPG are the most used unstructured data formats. Figure 3 shows the most commonly used formats.

The results show that HTML, PDF, CSV, XLS, and JPG are the most utilized data formats. Also, although comma-separated value (CSV) formats [51] are widely used, the use of RDF format is limited. Another finding is that proprietary formats such as DOC, XLS, XLSX, RAR, and AutoCAD, among others, are still used. Those data formats are not machine-readable [24]. Thus, those datasets cannot be used to enrich and add value to other data. On the other hand, some CKAN instances, such as Datahub.io (old version), provide additional nonproprietary formats for each of the published datasets, in order to apply open data principles.

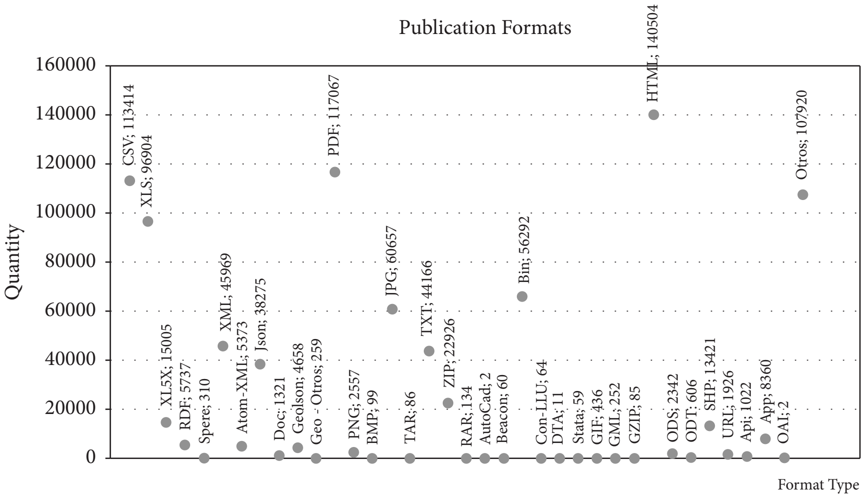

Figure 3: Format types.

The aim for three stars in the LOD model is a minimum requirement for open data publishing. However, licenses can be applied to data in any format (DOC, XLS, XLSX, RAR, AutoCAD, and among others), including those embedded within PDF documents. Some people use this kind of format due to the lack of a learning curve in machine-readable structured nonproprietary formats. As [2] said: "A proprietary file format is one that a company owns and controls. Data in this format may need proprietary software to be read reliably. Unlike an open format, the description of the format may be confidential or unpublished and can be changed by the company at any time. Proprietary software usually reads and saves data in its proprietary format. For example, different versions of Microsoft Excel use the proprietary XLS and XLSX formats."

Briefly, to reach the third level in the LOD model, the data are available using an open license in a widely reusable format, which means users do not need specific and proprietary software to reuse it [52].

On the other hand, although with a low level of use, different instances are using formats such as Atom, RDF, JSON, ODS, and SHP. Finally, in the sample of the 40 instances queried, we can identify other results: the Datahub instance holds the most extensive quantity of formats, the Rio Grande State Open Data Portal only shows data published in CSV (15,574 datasets published), the Salzburgerland Open Data Portal handles 13 datasets in SPARQL format, and lastly, 90% of the instances manage CSV as one of the data formats.

4.1.3. Data Updating. The timeliness principle argues that published datasets must be made available to the public promptly. As proposed in [16], data that are highly volatile should be up-to-date, and that is why priority should be given to time-sensitive data. Real-time information updates would maximize the usefulness that the public can get from this information. According to the findings, the highest proportion of datasets (60.5%) was updated less than six months ago. Figure 4 shows this monthly update distribution.

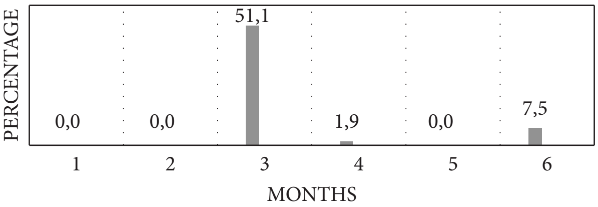

Figure 4: Percentages of datasets updated in the first six months.

However, results show that 13.9% of the datasets have not been updated between 2 and 4 years ago. As was described previously, the data quality dimension includes timeliness or currency [53]. It means that data have been updated to keep it current and are available to use when data are needed. According to the 10 principles of open data [54], datasets released should be made available to the public promptly (timeliness) whenever feasible, as quickly as it is gathered and collected, to preserve the value of the data. In short, we must be careful about the information we need. If time-sensitive datasets are not updated, data are not reliable and trustworthy. For that reason, we cannot ensure the accuracy and reliability of the data. On the other hand, some datasets have non-time-sensitive data, for example, historical population data. This information does not change over time, so it is reliable and trustworthy.

Another finding shows that those instances that handle the most quantity of datasets are those that have the most dispersed updating times of their datasets (Table 3). Furthermore, some datasets do not present an updating date, they have never been updated, or their update is given by updating one of their resources. Therefore, the lack of regular updating of those datasets that may change over time influences both the dataset quality and the search results performed by the consumer users.

4.1.4. Domains. As described by [5, 7, 55] and [25], datasets published on the Web can be classified by different knowledge domains such as Media, Publications, Life

Table 3: Instances with the most datasets.

Instance	0–6 months	6–12 months	12–18 months	18–24 months	24–48 months	48–60 months	More than 60 months	Total
dados_gov_br	4658	152	79	289	41	0	0	5219
data_bris_ac_ukdata	12914	0	0	0	0	0	0	12914
data_gov_uk	3012	2184	774	904	2587	1719	0	11180
data_overheid_nldata	12940	0	1	0	0	0	0	12941
datahub_io	0	174	322	708	3451	6605	3	11263
drdsi_jrc_ec_europa_eu	0	0	1105	3372	5562	171	0	10210
Open_alberta_ca	3798	3639	3000	680	2277	0	0	13394
data_go_jpdata	2444	2024	3872	1526	10329	0	0	20195
Europeandataportal	93812	7188	0	0	0	0	0	101000

Sciences, Geographic Data, User-Generated Content, Interdisciplinary, Government, Linguistics, Social Networking, Health, Education, and Environment. The domain tags identified in the 226,393 datasets queried are shown in Table 4.

Table 4: Domain tags by instance.

Instance	#Tags
Africa OpenData	2.453
Datos Gub uy	60
catalogue_data_wa_gov_au	1.613
dados_al_gov_br	1.208
dados_fortaleza_ce_gov_br	183
dados_gov_br	3.252
dados_recife_pe_gov_br	181
dados_rs_gov_br	196
dartportal_leeds_ac_uk	101
data_bris_ac_ukdata	901
data_gov_ie	3.431
data_gov_ro	838
data_gov_sken	370
data_gov_uk	9.111
data_kk_dk	530
data_lexingtonky_gov	0
data_nsw_gov_audata	861
data_ottawa_ca	191
www_hri_fi	676
zagreb_hr	16
data_overheid_nldata	6.139
datahub_cmap_illinois_gov	76
datahub_io	13.790
datapoa_com_br	8
datar_noip_me	161
datosabiertos_gob_ec	258
datosabiertos_malaga_eu	211
donnees_ville_montreal_qc_ca	847
drdsi_jrc_ec_europa_eu	23.335
ecaidata_org	112
gisdata_mn_gov	1.754
iatiregistry_org	251
open_alberta_ca	14.281
open_canada_cadata	0
opendata_aragon_esdatos	1.729
opendata_caceres_es	20
salzburgerland_com	13
surrey_ca	737
www_civicdata_io	64
www_data_go_jpdata	4.534
www_europeandataportal_eudata	64.651

Although the use of standard domain tags is identified (transport, health, policy, investment, statistics, geography, education, public sector, economy, and energy, for instance), we can identify a large number of domain tags that do not take account of a domain taxonomy.

Some causes of these multiple domains are the lack of guidelines about how to fill in the tag information, the lack of staff preparation who do this task, and the lack of support offered by the applications used to process this information. This disparity of domains makes it difficult to classify and process this type of information.

4.1.5. Registration of Authors and Organizations (Providers). Concerning the provenance information, results show that there is a wide dispersion in the provenance registrations (Table 5). More than half of the queried instances do not have provenance information fully registered, and eight of these queried instances do not have any provenance information registered in their datasets.

Finally, some of the results obtained are described in Table 6.

There are instances such as the State of Rio Grande and Lexington that do not report information about authorship or organizational provenance.

However, there are instances such as Salzburgerland, DART, and Montreal that publish complete authorship and provenance information about their datasets. In the main, despite the existence of tags and best practices for the provenance registration, the published datasets do not have this type of information or are handled half-finished, which influences the confidence assessment of the suppliers of the datasets operated.

4.2. The Ability to Reuse. For this perspective, a query interface that allows visualizing URLs used in each queried dataset was built. This visualization allowed us to analyze the linking subject or object behavior of each dataset. As a result, this study showed that queried datasets contain different types of resources, and each resource can be accessed using its URL. Japan's open data instance, which holds 20.195 datasets, manages 255.132 linked resources, for instance. The data instances with the highest number of linked resources are shown in Figure 5 and Table 2.

Table 5: List of authors and organizations of the queried instances.

No.	Open data instances/portal	Authors		Organizations	
		Record (%)	No. of records (%)	Record (%)	No. of records (%)
1	Ecuador	10.2	89.8	0.0	100.0
2	Datahub.io	52.8	47.2	54.5	45.5
3	Croatia	4.5	95.5	100.0	0.0
4	The State of Rio Grande, Brazil	0.0	100.0	0.0	100.0
5	The State of Alagoas, Brazil	86.0	14.0	50.3	49.7
6	The African Continent	29.0	71.0	14.0	86.0
7	Malaga	98.0	2.0	100.0	0.0
8	UK Open Data	12.0	88.0	48.0	52.0
9	Salzburgerland	100.0	0.0	100.0	0.0
10	Cáceres	100.0	0.0	0.0	100.0
11	The University of Bristol	98.8	1.2	100.0	0.0
12	European Data Portal	0.0	100.0	99.9	0.1
13	Puerto Alegre, Brazil	0.0	100.0	100.0	0.0
14	New South Wales Government Public Data	27.6	72.4	40.2	59.8
15	Alberta, Canada	8.1	91.9	3.8	96.2
16	Ottawa, Canada	95.4	4.6	0.0	100.0
17	The Prefecture of Recife, Brazil	96.7	3.3	98.4	1.6
18	Surrey, Canada	4.7	95.3	100.0	0.0
19	Copenhagen	69.1	30.9	93.8	6.2
20	Montevideo	93.4	6.6	92.9	7.1
21	Joint Research Centre (JRC) of the European Commission	0.0	100.0	44.5	55.5
22	Lexington	0.0	100.0	0.0	100.0
23	Helsinki	100.0	0.0	87.6	12.4
24	Brazilian	95.3	4.7	97.4	2.6
25	Unofficial repository of open public data of the Argentine Republic	51.9	48.1	43.4	56.6
26	Ireland	14.9	85.1	2.3	97.7
27	Data Portal Project Detection of Archaeological Residues (DART)	100.0	0.0	100.0	0.0
28	Western Australia Government	86.3	13.7	99.9	0.1
29	Romania	29.8	70.2	4.0	96.0
30	Fortaleza, Brazil	3.2	96.8	100.0	0.0
31	Slovenia	96.0	4.0	94.3	5.7
32	Montreal, Canada	100.0	0.0	100.0	0.0
33	Dutch Government Data Portal	47.9	52.1	11.8	88.2
34	Aragon	97.6	2.4	88.6	11.4
35	Chicago Metropolitan Agency for Planning Data Center	99.1	0.9	98.7	1.3
36	The International AID Transparency Initiative	2.1	97.9	0.0	100.0
37	ECAI	0.1	99.9	99.2	0.8
38	CivicData	3.2	96.8	5.8	94.2
39	Minnesota	0.0	100.0	100.0	0.0
40	Japan	0.0	100.0	100.0	0.0

Table 6: Authorship and organization registry.

Percentage of instances	Description
20	It does not register any data of dataset authorship
15	It does not register any data of the organization that provides the dataset
12.5	These instances register 100% of the authoring information of their dataset
27.5	These instances register 100% of the provenance organization information of their dataset

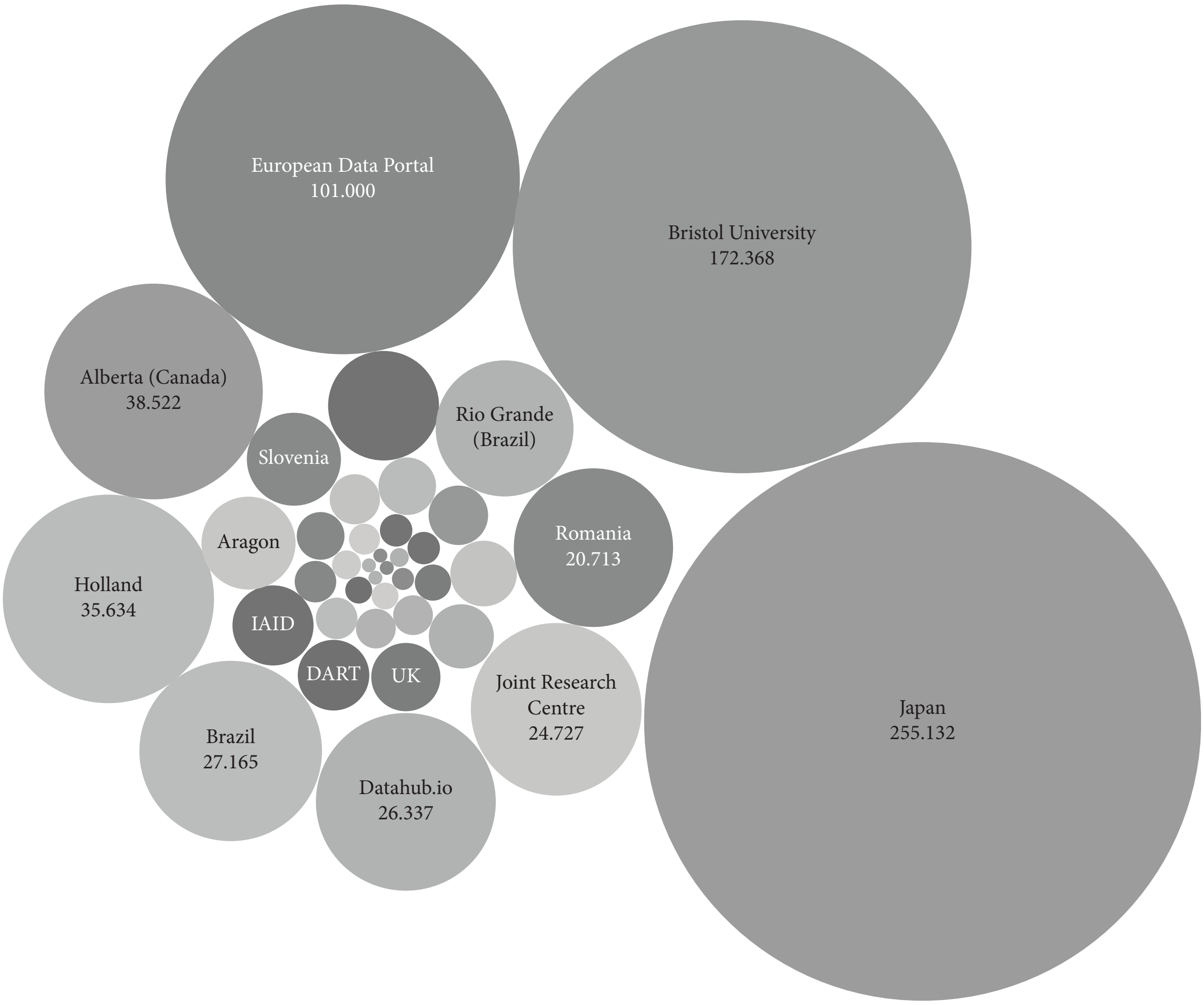

Figure 5: Number of linked resources.

These results let us identify problems such as restrictive license types, the lack of licensing definition, and reduced use of structured and nonproprietary formats. Despite these issues, organizations that publish their data in CKAN instances use active URLs. However, in some cases, these URLs link proprietary or no-structured files. The dataset of the ongoing recruitments of the Municipality of Lorca (European Data Portal) is an example of this issue. This instance links some files in Excel, which do not load or display information.

On the contrary, it is necessary to highlight the work and evolution that CKAN has been providing, to improve the services of publication and consumption of data. Concerning its internal structure, datasets have tags to describe different aspects. One of these aspects is the description of its behavior as a linkage subject or object to other datasets (relationships_as_object and relationships_as_subject). When looking into the instances queried, only 2 of the 40 of them provide information about their behavior as a linkage subject or object to other datasets: Datahub.io (170 subject-object tags) and The University of Bristol (12437 subject-object tags). It shows that, even when resources provide active and reachable URLs, the tagging structure does not provide complete information. Only those organizations that both create publication services and decide to generate data consumption start to review this type of small details, which provide primary information at the time of analyzing the linking level of datasets published on the Web.

In a nutshell, although efforts made in openness and linked data are shown, the low metadata quality and the weak application of the best practices of availability and reuse have created barriers that discourage the growth and use of the Web of Data. Although nonproprietary formats and reachable URLs [56] are used, data-publishing problems were identified. Issues, such as the nonupdated datasets, the diversity of published formats, the nonproper filling out tags, the low availability of end-user-friendly tools, the poor institutional policy oriented to linked data, and the lack of guidelines for data providers, were identified. These problems reduce the use of the Web of Data. Considering these factors, platforms like Datahub have opted for a new orientation called Frictionless Data [57]. This strategy provides a simple wrapper and basic structure for data transport, significantly reducing friction in data exchange and integration and also supporting automation without imposing significant changes on the underlying data being packaged.

5. Discussion of Results

First of all, concerning the research findings, it is seen that in the Government domain, the Web of Data suffers from a set of issues in different implementation stages. In the case of open resources availability (levels 1 to 3 of LOD schema), the results show that although efforts are made to publish data (level 1), these processes present different barriers such as the use of proprietary formats, the multiplicity of data formats, the outdated nature of the published data, the lack of appropriate licensing allowing the use, reuse, and distribution of published data. In the case of the ability to reuse, published datasets and query services are identified in the queried instances. Likewise, queried datasets have URIs that link information to their triples. However, the behavior of linking subject or object is not recorded in the datasets. This problem is due to human errors or the absence of knowledge of the dataset structure.

Despite datasets being available in the queried instances, the following findings were identified:

(i) A portion of datasets do not have a license, or it is too specific. The lack of a copyright declaration does not allow the reuse of the data and restricts checking the data attribution and share-alike requirements.

(ii) The CSV format is one of the most used formats in the queried instances. However, the lack of RDF formats reduces the scope of the linked data.

(iii) Some datasets have been updated six months before. However, some datasets were updated more than two years ago. This variation of updates affects the data timeliness for reuse and decision-making based on its content.

(iv) As far as domain tags are concerned, a significant disparity of domain names and problems in their tagging is identified. Different abstractions and particular designs of real-world objects can generate these problems.

(v) A large proportion of datasets do not record provenance information. The lack of relevant information makes it difficult to determine whether the dataset fits the purpose of information required by the user, affecting the user's confidence.

(vi) Although resources are linked using their URLs, this information is not recorded in the dataset structure. This lack of registration may be due to human errors or a lack of detailed knowledge of the data structure.

(vii) The government information changes from one dataset to another because of different abstractions of the knowledge domain, in addition to a limited expressivity of the used vocabulary.

Considering that metadata published reveals different data quality problems, these findings reinforce the hypothesis of the decline that the Web of Data presents. These problems have been identified by authors in [17–21, 58] and [59], among others, which are detrimental both to the dataset information and to the information obtained from the queries. Data quality is crucial when it comes to making far-reaching decisions based on the results of querying multiple datasets [16].

While there are lots of open data guidelines, putting it into practice, it is a little bit difficult. In terms of data availability (levels 1–3 of linked data), we can identify stumbling points as follows.

Some people think that open data merely requires that each opening proposal includes files published on the Web, and no specific practices are specified since best practices are often different for different projects. Simply making data publish does not guarantee that the data have utility as open data. A substantial proportion of published data is not available under an open licensing, had insufficient metadata, uses an inappropriate file format, or is out-of-date. Briefly, openness is changing to what can be termed "open-washing" [60] which means data are open but are not complete or there are data qualities and discovery issues.

While there are a lot of open data standards and best practices, most people are not familiar with them and do not use them. Opening practices are decentralized, and people in charge of data opening rarely receive formal training about opening and linked data. They are often left to abstract and represent their data models. Our results show that it is necessary to improve the first stages in the linked data lifecycle (abstraction, modeling, and opening). We have to shift our paradigm from "open-washing" to "opening focused on data availability."

In terms of data reuse, according to levels 4 and 5 of linked data, we can identify stumbling points as follows.

Some people think that their datasets merely require that each metadata adds some kind of literal data and does not use information from outside their immediate environment. On the other hand, an inappropriate open data license does not allow our data to use by other people. Lastly, people do not concern about the visibility of their data as they think that they alone shall have the right to use them.

We have to shift our paradigm towards the reuse of open data, understanding that semantic enrichment allows that any metadata inside our dataset can be enriched by information from outside our immediate environment, which we reused. For this purpose, it is essential to both share our dataset identification properly and link other datasets using their URIs. Additionally, it is necessary to record this information inside the RDF structure of our datasets.

Keeping in mind that data availability is the base for data reuse, we have to use best practices in opening and linking data in the first place. On the other side, data owners must shift their paradigm, and they need to understand that open data, as a social movement, will become part of their workflow and understand that this process requires policies, investment, and training. The necessity of open, share, and reuse data has never been more apparent than it is today when we are all suffering to some degree from the COVID-19 crisis.

Lastly, although existing successful Web of Data examples, such as Place Name Databases, where you can find open data about place names, or Census 2006 as linked open data, a project developed by the Irish Central Statistics Office [61]; various organizations do not understand the purpose of publishing data on the Web, let alone why data on the Web should be linked [61–63]. For that reason, we have to understand that data enrichment is not an exclusive task of the "public sector." There are lots of data enrichment experiences in sectors that do not make their data available to the public.

In general, our findings can be identified in studies as follows.

Data availability and reuse issues permeate different knowledge domains. An example of which is the digital humanities researchers. In the Arts and Humanities, this tension between publishing research results in silos, particularly where the underlying research data are never shared, and the desire for open, reusable data remains [64]. In addition, a main obstacle to the reuse of digitized cultural heritage is a lack of data quality. Reusing LOD datasets is a challenging task requiring the knowledge of several technologies as well as how the data are modeled [65]. On the other hand, regardless of the specific tasks that LOD-based tools aim to address, the reuse of such knowledge may be challenging for diverse reasons, e.g., semantic heterogeneity, provenance, and data quality [66]. Another problem potentially is the inability of machines to automatically find and read data, which makes it challenging for the data to be reused by any stakeholder. Thus, if the data are not in some way open or accessible, it is impossible to reuse the data for other purposes, like AI [67].

In the knowledge domain of linguistics, there is a need to share linguistic resources, but reuse is impaired by several constraints including a lack of common formats, differences in conceptual notions, and unsystematic metadata. The following five constraints are discerned in this knowledge domain [68]:

(1) Linguistic resources are often designed for particular tasks (e.g., part-of-speech tagging, named entity recognition).

(2) There is a plethora of different markup languages, which are often not fully compatible between systems, much less between domains.

(3) Each linguistic resource may use different conceptual models. For example, there are dozens of different part-of-speech tag sets.

(4) Existing linguistic resources often do not provide precise or machine-readable definitions of the terminology they use, thus making it difficult to reuse them without manual investigation.

(5) It is often difficult to obtain the full metadata around the creation of a resource.

Briefly, if you want to start with LOD, keep in mind the recommendation of [69]: "Many organizations are interested in publishing linked open data. However, this is a complex endeavor that requires a gradual approach, especially in situations where resources are scarce and technical know-how and infrastructure need to be developed first. In such contexts, it is recommended that organizations 'open data first, and then link' (Caracciolo & Keizer 2015), focusing on priority datasets that are highly visible or which have high reuse value."

The studies described previously let us identify that our results fit into the broader research context on resource availability and ability to reuse. On the other hand, these studies evidence that the sample used in our research, taken from different countries and topics, is representative given that it lets us evaluate and identify that the availability and reusability issues are present nowadays and need to be addressed to improve the data quality. The lack of appropriate metadata limits data openness and enrichment. Using metadata with the correct metadata architecture can yield considerable benefits for LOD publication and use, including improving finding ability, accessibility, storing, preservation, analyzing, comparing, reproducing, finding inconsistencies, correct interpretation, visualizing, linking data, assessing and ranking the quality of data, and avoiding unnecessary duplication of data [70].

6. Conclusion and Future Work

The Semantic Web has faced challenges that have emphasized aspects that have not allowed its evolution at the expected pace. These factors have significant features associated with open data that need to be evaluated: Are datasets displayed in machine-readable formats? Are they reusable? Are they free of charge? Do they have an open license? Are they up-to-date? Is it easy to get information about them? These features, among others, are still challenges that affect the availability and reusability of data within the Semantic Web. Once the study is completed, the results obtained from the dataset exploitation let us identify the following:

(i) Datasets are neither representing abstractions that respond to the same context nor being described with known vocabularies. Additionally, factors, such as the lack of updating, the poor technological support, and the prominent learning curve, are barriers to the development of linked data projects. Finally, the lack of knowledge about restrictions and permissions acquired on the data restricts access to them, thereby limiting consumers' ability to exploit the possibilities of open data.

(ii) Regarding the reusability of linked resources, most of the queried datasets make use of URIs to connect the information to their triples. However, a low rate of the queried instances records their behavior as a linkage subject or object. Although the absence of these registers does not directly affect the operation of linked data, such problems may be due to a lack of awareness of RDF structure or human errors.

Based on these research results, people may feel those open data portals are good enough for a political party, a public institution, or the Government but not good enough for linked data. This assertion can be "right" because data are appropriate for their target audience. But this assertion is not correct. Government data portals that do not meet the minimum requirements for opening their data can be identified. Data portals must provide a good level of open data quality for data openness and linking processes. Besides, the linked resources suffer from abstraction and data quality issues despite the existence of best practices for data publication. In addition, although the linked resources use reachable URIs, the registration of linkages is a major issue when you try to enrich your data. For that reason, the link analysis is proposed as a strategy to complement the reuse analysis of the linked resources.

Regarding availability and reuse approaches, the research results evidence that these approaches impact the data published on the Web. These approaches are indispensable components to reach the first levels of the linked data model and allow evaluation of the data access by the users, who are looking to connect their data and improve it with timely and accurate data.

Although this research analyzed the governmental domain, these variables can be worked out in other knowledge domains because they represent operational aspects that fit any knowledge domain. Besides, this method is applicable to scientific data portals as long as their metadata are published in an open format and the platform works over a CKAN instance. However, the description of the information in the RDF model must be a thoughtfully completed task to improve the interoperability levels in the knowledge domain worked.

Concerning the research findings, the following issues are identified: (a) the lack of update that restricts the timeliness of the data, (b) the lack of licensing skews the use of the data, (c) the absence of information that allows determining if the dataset suits the purpose of information required by the user, (d) difficulties to access and availability which does not permit the data exploitation, and (e) the discrepancy of real-world abstractions reduces the interoperability between repositories. All of this evidence has been weakening the linked resources perspective, shifting towards strategies that allow data transport without needing their prior processing.

Last but not least, based on the results of this study, the challenges that we have identified as the main ones to address in the information linking process, from the data availability and reuse approaches, are the following:

(i) Open and link data as a priority in organizations (at all levels of organizations). This organizational priority will offer many business opportunities to consumer users. Furthermore, enough trained human resources and the necessary budget must be provided to carry out the task of open and linked data. Also, the open data legislation should be provided, which extends to different government and organization levels, whether public or private.

(ii) To empower users, increasing their perception of the possibilities of reuse of information, in addition to providing them with agile and straightforward tools and procedures to carry out the open and linked data processes.

(iii) The technological context may be improved through mechanisms that easing the tasks of publishing and linking data, managing the metadata of published resources. Also, these tools must provide stable services that allow users to exploit data in a more user-friendly way. Finally, the learning curve in topics of standards, vocabulary, and languages used for the knowledge representation must be improved.

(iv) To define and unify the data publication and linking workflows properly. The proper definition and organization of activities for opening, publishing, and enriching resources help streamline the development of tasks (curation, acquisitions, discovery, and analytics) and facilitate coordination among people. Besides, planning activities in the openness and enrichment resource process allows for identifying needs in the open and linked data learning curve. Finally, this workflow must be easy to address by data editors and consumer users.

(v) A well-defined process for abstracting real-world entities and attributes must be established in order to improve interoperability between repositories. Both the proper knowledge domain contextualization and the use of vocabularies with enough expressiveness are vital for data modeling workflow.

(vi) To complement availability and reuse features, RDF data formats that meet all LOD model requirements should be provided.

(vii) Linked data should provide strategies to overcome not only the current difficulties of linking but also the search for resources that contribute to data enrichment.

(viii) Given that in different researches, the linked data quality is mainly evaluated on the generated instances, and it is necessary to strengthen the abstraction and design of data models, based on linked data quality requirements.

(ix) Also, given the particular designs of real-world objects combined with the failure to open data models and the use of their standards, it is necessary to move towards the opening and the reusing, safeguarding the fundamental rights. Moreover, the idea should be to decrease the syntactic (languages) and semantic (meanings) transformations that ought to be carried out to use the data.

(x) Licensing and copyright: the announcement of publishing rights and the respective linked data authorization allow to both reuses and the use of the data legally by complying with the restrictions provided for its manipulation.

(xi) Design and implementation of data update policies and strategies to leverage appropriate data.

Some practical recommendations for organizations looking to implement open and linked data could be to increase the stakeholder's learning curve on applying LOD principles and improving metadata quality on the resource descriptions. Besides, organizations must understand that opening and enriching their data require a completely new approach, and they have to pay special attention and control to this project, generally by putting money, the commitment by management at all levels, and lots of time. And last, organizations must apply the open data and linked open data principles to their published dataset to add real value to their data.

Implementing explicit concepts and rules abstraction needed to build particular models in a specific domain is proposed as future work. This proposal raises the design of a metamodel that allows the generation of data instances based on quality dimensions of linked data and avoids differences in the context represented. Other topics, such as investigating the effectiveness of different strategies to improve data quality, evaluating the impact of open data legislation, or exploring the potential of new technologies such as blockchain or artificial intelligence for linked data, are proposed as future works.

Data Availability

The data used are as appendices in the article, there are a total of 4 Appendix, which has the data used.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This study has been developed from the framework of the doctoral research project in Linked Data at the Universidad Distrital Francisco José de Caldas. Similarly, work is being done on this issue as a working line of the GIIRA Research Group.

References

[1] F. Bauer and M. Kaltenböck, "Linked open data: the essentials," *First Edition. Edition mono/monochrom*, vol. 710, 2017.

[2] Oki, *Open Data Handbook*, Open Knowledge International, Cambridge, United Kingdom, 2019.

[3] T. Berners-Lee, T. Heath, and C. Bizer, "Linked data-the story so far," in *Semantic services, interoperability and web applications: emerging concepts*, IGI global, Pennsylvania, PL, USA, 2009.

[4] A. Abella, M. Ortiz-de-Urbina-Criado, and C. De-Pablos-Heredero, "Indicadores de calidad de datos abiertos: el caso del portal de datos abiertos de Barcelona," *El Profesional de la Información*, vol. 27, no. 2, pp. 375–382, 2018.

[5] T. Heath and C. Bizer, *Linked Data: Evolving the Web into a Global Data Space*, Morgan & Claypool, San Rafael, California USA, 2011.

[6] State of Lod Cloud, "Lod2 statistical office workbench," 2019, http://lod2.stat.gov.rs/lod2statworkbench.

[7] J. McCrae, *The Linked Open Data Cloud*, Insight Centre for Data Analytics, Belfield, Dublin 4, Ireland, 2020.

[8] Ckan, *Ckan Api Guide*, Open Knowledge International, Cambridge, United Kingdom, 2018.

[9] J. F. Herrera-Cubides, P. A. Gaona-García, J. Alonso-Echeverri, K. Riaño-Vargas, and A. Gómez-Acosta, "A fuzzy logic system to evaluate levels of trust on linked open data resources," *Revista Facultad de Ingeniería, Issue*, vol. 86, pp. 40–53, 2018.

[10] R. W. Ckan, "Guía de usuario," 2013, https://github.com/ckan/ckan/wiki/Gu.C3.ADa-de-usuario.

[11] O. D. Open, *Definition 2.1*, Open Knowledge International, Cambridge, United Kingdom, 2019.

[12] J. Tauberer, "The annotated 8 principles of open government data," 2007, https://opengovdata.org/.

[13] J. Tauberer, *Open Government Data The Book*, Perfect-bound Paperback, New Hampshire, United Kingdom, 2014.

[14] S. Auer, J. Lehmann, and A. Ngomo, "Introduction to linked data and its lifecycle on the web. Semantic technologies for the web of data reasoning Web," *Lecture Notes in Computer Science*, vol. 6848, 2011.

[15] M. Morsey, J. Lehmann, S. Auer, C. Stadler, and S. Hellmann, "DBpedia and the live extraction of structured data from wikipedia," *Electronic library and information systems*, vol. 46, pp. 157–181, 2012.

[16] A. Zaveri, A. Rula, A. Maurino, R. Pietrobon, J. Lehmann, and S. Auer, "Quality assessment methodologies for linked open data. A systematic literature review and conceptual framework," *Semantic Web Journal*, vol. 1, pp. 1–33, 2012.

[17] B. Farias-Lóscio, C. Burle, and N. Calegari, "Data on the web best practices," 2019, https://w3c.github.io/dwbp/bp.html.

[18] B. Behkamal, M. Kahani, E. Bagheri, and Z. Jeremic, "A metrics-driven approach for quality assessment of linked open data," *Journal of Theoretical and Applied Electronic Commerce Research*, vol. 9, no. 2, pp. 64–79, 2014.

[19] P. Király, "A metadata quality assurance framework," in *Proceedings of the 8th International Conference on Qualitative*

and Quantitative Methods in Libraries, Gottingen, Germany, September 2016.

[20] P. Mendes, “Conceptual model and best practices for high-quality metadata publishing. planet data,” 2012, https://docplayer.net/40848695-D2-1-conceptual-model-and-best-practices-for-high-quality-metadata-publishing.html.

[21] E. R. M. Vidal, S. Castillo, O. Burguillos, and O. Baldizan, “Analyzing linked data quality with liquate. the semantic web: eswc 2014 satellite events eswc,” *Lecture Notes in Computer Science*, vol. 8798, pp. 488–493, 2014.

[22] G. Karvounarakis, I. Fundulaki, and V. Christophides, “Provenance for linked data. In search of elegance in the theory and practice of computation,” *Lecture Notes in Computer Science*, vol. 8000, 2013.

[23] N. Fayyaz, I. Ullah, and S. Khusro, “On the current state of linked open data: issues, challenges, and future directions,” *International Journal on Semantic Web and Information Systems*, vol. 14, no. 4, pp. 110–128, 2018.

[24] J. Tauberer, *Ten Principles for Opening up Government Information*, Sunlight Foundation, Washington, D.C., USA, 2017.

[25] World Wide Web Foundation, *The Open Data Barometer*, World Wide Web Foundation, Washington, D.C, USA, 2019.

[26] O. Can and D. Yilmazer, “A novel approach to provenance management for privacy preservation,” *Journal of Information Science*, vol. 46, no. 2, pp. 147–160, 2020.

[27] Experian, “Global Data Management Research. Taking Control in the Digital Age,” Experian Tech, USA, Benchmark Report, 2019.

[28] R. Vaidyambath, J. Debattista, N. Srivatsa, and R. Brennan, “An intelligent linked data quality dashboard,” in *Proceedings of the AICS 27th AIAI Irish Conference on Artificial Intelligence and Cognitive Science*, Galway, Ireland, December 2019.

[29] J. D. Fernández, M. A. Martínez-Prieto, P. de la Fuente, and C. Gutiérrez, “Characterizing RDF data sets,” *Journal of Information Science*, vol. 44, no. 2, pp. 203–229, 2018.

[30] S. Dietze, H. Yu, D. Giordano, E. Kaldoudi, N. Dovrolis, and D. Taibi, “Linked education: interlinking educational resources and the Web of Data,” in *Proceedings of the 27th ACM Symposium on Applied Computing (SAC-2012)*, Trento, Italy, March 2012.

[31] Godi, “The global open data index,” 2019, https://index.okfn.org/.

[32] J. F. Herrera-Cubides, P. A. Gaona-García, and K. Gordillo-Orjuela, “A view of the web of data,” *Case Study: Use of Services CKAN. Ingeniería. Universidad Distrital Francisco José de Caldas*, vol. 22, no. 1, pp. 111–124, 2017.

[33] J. F. Herrera-Cubides, P. A. Gaona-Garcia, and S. Sánchez-Alonso, “The web of data: past, present and ¿future?” in *Proceedings of the XI Latin American Conference on Learning Objects and Technology (LACLO)*, pp. 1–8, San Carlos, CA, USA, October 2016.

[34] A. Hogan, J. Umbrich, A. Harth, R. Cyganiak, A. Polleres, and S. Decker, “An empirical survey of Linked Data conformance,” *Journal of Web Semantics Science Direct*, vol. 14, pp. 14–44, 2012.

[35] J. Klímek, P. Skoda, and M. Necaský, “Survey of tools for linked data consumption,” *Czech Republic, Faculty of Mathematics and Physics*, Charles University, Staré Město, Czechiapp. 1–57, 2018.

[36] K. Smith-Yoshimura, “Linked data implementations: who, what, why?” in *Semantic Web in Libraries SWIB18 Semantic Web in Libraries*, Bonn, Germany, 2018.

[37] European-Union, “Portal europeo de datos,” 2019, https://www.europeandataportal.eu/es.

[38] M. Schmachtenberg, C. Bizer, and H. Paulheim, “Adoption of the linked data best practices in different topical domains,” in *Proceedings of the Semantic Web–ISWC 2014. ISWC 2014. Lecture Notes in Computer Science*, Riva del Garda, Italy, October 2014.

[39] Ontotext, “What are linked data and linked open data?,” 2023, https://www.ontotext.com/knowledgehub/fundamentals/linked-data-linked-open-data/.

[40] Stardata, “Star open data,” 2015, https://5stardata.info/en/.

[41] J. Riley, *Understanding Metadata what Is Metadata, and what Is it for*, NISO Primer Series, Baltimore, MD, 2017.

[42] D. Wood, M. Zaidman, L. Ruth, and M. Hausenblas, *Linked Data Structured Data on the Web*, Manning Publications, Shelter Island, NY, USA, 2014.

[43] A. Assaf, R. Troncy, and A. Senart, “Hdl-towards a harmonized dataset model for open data portals,” 2015, http://www.eurecom.fr/%7Etroncy/Publications/Assaf_Troncy-profiles15b.pdf.

[44] L. Thomas, “Simple random sampling | definition, steps & examples,” 2023, https://www.scribbr.com/methodology/simple-random-sampling.

[45] A. Hayes, “Simple random sampling: 6 basic steps with examples,” 2023, https://www.investopedia.com/terms/s/simple-random-sample.asp.

[46] S. Digital, “What is CKAN?,” 2023, https://salsa.digital/insights/what-is-ckan.

[47] BuiltWith, “Websites using ckan,” 2023, https://trends.builtwith.com/websitelist/CKAN.

[48] A. Varón-Capera, P. A. Gaona-García, J. F. Herrera-Cubides, and C. Montenegro-Marín, “VACIT tool for consumption, analysis and machine learning for LOD resources on CKAN instances,” *Information systems and technologies to support learning. EMENA-ISTL 2018. Smart Innovation, Systems and Technologies*, vol. 111, pp. 552–564, 2018.

[49] O. D. Json, “Format (odata version 2.0),” 2018, https://www.odata.org/documentation/odata-version-2-0/json-format/.

[50] L. Ding and T. Finin, “Characterizing the semantic web on the web. The semantic web-ISWC 2006,” *Lecture Notes in Computer Science*, vol. 4273, pp. 242–257, 2006.

[51] Web Working Group, “All standards and drafts: csv on the web a primer,” 2019, https://www.w3.org/TR/?tag=data#w3c_all.

[52] A. Gray, “Tpximpact. understanding linked data principles,” 2023, https://www.tpximpact.com/knowledge-hub/blogs/tech/linked-data-principles/.

[53] P. Connor, “Open data 101: the history and principles of open data,” 2023, https://apolitical.co/solution-articles/en/open-data-101-the-history-and-principles-of-open-data-part-1.

[54] OpenGovData, “The 8 principles of open government data,” 2007, https://opengovdata.org/.

[55] C. Bizer, “The emerging web of linked data,” *IEEE Intelligent Systems*, vol. 24, no. 5, pp. 87–92, 2009.

[56] E. Rajabi, S. Sanchez-Alonso, and M. Sicilia, “Analyzing broken links on the Web of Data an experiment with DBpedia,” *Journal of the American Society for Information Science and Technology*, vol. 65, no. 8, pp. 1721–1727, 2014.

[57] Open Knowledge International, *Frictionless Data. Frictionless Data, Specifications and Software*, Open Knowledge International, Cambridge, United Kingdom, 2019.

[58] Igi, *Information Retrieval and Management: Concepts, Methodologies, Tools, and Applications*, IGI Global, Information Resources Management Association, Hershey, Pennsylvania, USA, 2018.

[59] A. Zaveri and A. Rula, "Methodology for assessment of linked data quality ldq 2014-1st workshop on linked data quality," 2014, http://ceur-ws.org/Vol-1215/paper-04.pdf.

[60] C. Villum, *Open-washing–the Difference between Opening Your Data and Simply Making Them Available*, Open Knowledge Foundation, Cambridge, United Kingdom, 2014.

[61] W. Derguech and E. MacCuirc, "The irish experience in publishing linked open data. european data portal–linked data workshop," 2018, https://data.gov.ie/blog/lod-edp-dec-18.

[62] E. MacCuirc, *Open Data Is Coming: The Irish Experience. Databank and Dissemination*, Central Statistics Office, Ireland, 2018.

[63] Ece, "Guidance on common elements of statistical legislation," in *Proceedings of the Economic Commission for Europe. Conference of European Statisticians*, Geneva, Switzerland, June 2018.

[64] Hdh, "Linked open data in the arts and humanities," 2022, https://humanidadesdigitaleshispanicas.es/ijhac-a-journal-of-digital-humanities/.

[65] G. Candela, "An automatic data quality approach to assess semantic data from cultural heritage institutions," *Journal of the Association for Information Science and Technology*, vol. 74, 2023.

[66] V. Presuti, "Linked open data validity-a technical report from isws 2018," 2019, https://arxiv.org/abs/1903.12554.

[67] P. H. P. Jati, Y. Lin, S. Nodehi, D. Cahyono, and M. Van Reisen, "Fair versus open data: a comparison of objectives and principles," *Data Intelligence*, vol. 4, no. 4, pp. 867–881, 2022.

[68] M. Van Erp, "Reusing linguistic resources: tasks and goals for a linked data approach," in *Linked Data in Linguistics*, C. Chiarcos, S. Nordhoff, and S. Hellmann, Eds., Springer, Berlin, Heidelberg, 2012.

[69] L. Gonzalez, "Chapter 5 linked open data," 2018, https://unstats.un.org/wiki/pages/viewpage.action?pageId=36144023.

[70] A. Zuiderwijk, K. Jeffery, and M. Janssen, "The potential of metadata for linked open data and its value for users and publishers," *JeDEM-eJournal of eDemocracy and Open Government*, vol. 4, no. 2, pp. 222–244, 2012.