- ORIGINAL ARTICLE -

Publication of Linked Open Data – A Systematic Literature Review for Identifying Problems and Technical Tools Supporting the Process

Publicación de Datos Abiertos Enlazados: Revisión Sistemática de la Literatura para Identificar Problemas y Herramientas Técnicas de Apoyo al Proceso

Jairo H. Silva-Aguilar[1], Rommel Torres T.[2] and Elsa Estevez[3,4]

[1] *Facultad de Informática, Universidad Nacional de la Plata, Argentina*
[2] *Universidad Técnica Particular de Loja, Ecuador*
[3] *Departamento de Ciencias e Ingeniería de la Computación, Universidad Nacional del Sur (UNS), Argentina*
[4] *Instituto de Ciencias e Ingeniería de la Computación (UNS-CONICET), Argentina*
jairosilva@hotmail.com, rovitor@utpl.edu.ec, ecestevez@gmail.com

Abstract

On the Internet, we find a large amount of information from government institutions that has been published in open format. However, only a part of these data is available in standard formats such as Resource Description Framework (RDF), and to a lesser extent, is published as Linked Open Data (LOD). The main objective of the research presented in this paper is to identify problems and tools used in the process of publishing LOD with the purpose of establishing a basis for the construction of a future framework that will help public institutions to facilitate such processes. To fulfill the objective, we conducted a systematic literature review in order to assess the state-of-the-art in this matter. The contribution of this work is to identify the frequent problems that arise in the LOD publishing process. It also provides a detail of the frameworks proposed in scientific papers grouping the technical tools by phases that correspond to the LOD publication life cycle. In addition, it compiles the characteristics of the ETL (Extract-Transform-Load) tools that predominate in this review, such as Pentaho Data Integration (Kettle) and OpenRefine.

Keywords: Framework, ICT Tools, Linked Open Data, Open Data, Open Government.

Resumen

En Internet encontramos una gran cantidad de información procedente de instituciones de gobierno que se ha publicado en formato abierto. Sin embargo, sólo una parte de estos datos está disponible en formatos estándar como Resource Description Framework (RDF) y, en menor medida, se publican como Linked Open Data (LOD). El objetivo principal de la investigación presentada en esta publicación es identificar los problemas y las herramientas utilizadas en el proceso de publicación de LOD con el fin de establecer una base para la construcción de un futuro marco que ayude a las instituciones públicas a facilitar dichos procesos. Para cumplir con el objetivo, realizamos una revisión sistemática de la literatura con el fin de evaluar el estado del arte en la materia. La contribución de este trabajo es identificar los problemas frecuentes que surgen en el proceso de publicación de LOD. También proporciona un detalle de los marcos propuestos en artículos científicos agrupando las herramientas técnicas por fases que corresponden al ciclo de vida de la publicación LOD. Además, recopila las características de las herramientas ETL (Extract-Transform-Load) que predominan en esta revisión, como Pentaho Data Integration (Kettle) y OpenRefine.

Palabras claves: Marco de Trabajo, Herramientas TIC, Datos Abiertos Enlazados, Datos Abiertos, Gobierno Abierto.

1. Introduction

Open Government Data (OGD) literature explains that there are several ways in which institutions may appear transparent, but maintain certain opaque practices, for example, publishing data that are not understood by users, not publishing enough data, or publishing only data that an organization considers safe for them to publish [1]. An open government must interact with society in a transparent and participatory way, for which they need to publish data generated by public institutions, allowing its reuse.

Among others, the impact of being able to access and manipulate public data leads, in some cases, for actors of the ecosystem to collaborate in economic, scientific, administrative, and other tasks inherent to the public sector. Based on the above, a government would have two main reasons to publish open data. The first refers to the development of democratic values by giving access to data and consequently promoting transparency. The second deals with the development of the economy that allows the growth of information and its added social value [2,3].

The large amount of unstructured information available on the Internet has motivated the use and improvement of web semantic techniques to provide in the public domain more linked and meaningful information. In turn, this raises the need to provide tools and strategies aimed at the publication and access to data on the web. As a result, the concept of Linked Data (LD) was adopted, and currently, much of the information on the web is represented using LD formats such as RDF. However, data on the Internet is useful when there are no restrictions on their use, which refers to the concept of LOD [4].

In the case of Ecuador, we have reviewed several sets of published data and found that there is a large amount of heterogeneous data and in most cases, the evaluation of the quality of such data does not exceed the score of three stars in the system proposed by Berners-Lee [5]. It is rare to find data properly structured with Uniform Resource Identifier (URI), in RDF formats or linked to other similar datasets. According to Abida et al. [6], there is a lack of integrated solutions with automatic support to combine the right tools to help users, especially non-experts, to efficiently manipulate datasets, from extraction to publication, and to track progress during this process, as well as a semantic visualization of their data.

Given this scenario and the identified gap, our work focuses on a systematic literature review of the LOD publication life cycle that leads us to summarize the technical tools that are being used at each stage, as well as to expose the problems that exist during the process. The relevance of this research work focuses on the importance of knowing the technical tools that are used to support the publication of LOD and, based on this, to develop a framework to help professionals in public institutions in conducting such processes, taking into consideration that our goal is that the practical application of the research contributes to open government in its aim to improving transparency and citizen participation.

The rest of the document is organized as follows. After this Introduction, Section 2 details the methodology used for conducting the research work. Section 3 presents the results obtained from the literature review, which are discussed in Section 4. Finally, Section 5 presents the conclusions.

2. Methodology

A systematic literature review is considered a methodology to identify, evaluate and interpret relevant information in response to a well-defined research question, by collecting studies evaluated with quality criteria in a delimited period of time [7]. In this work, we have adopted the guidelines for conducting systematic reviews proposed by Kitchenham and Charters [8], which is summarized in three main phases: Planning, Conducting, and Reporting the Review. In each phase, we executed several tasks, which we explain below.

2.1. Planning the Review

To identify the need for a literature review, we noted that among the open data published on the web, as far as public entities are concerned, there are several data formats such as csv, pdf, Excel sheets, xls, xml, rdf, and json. RDF is a format that can be processed by a computer and also serves to exchange data between applications while preserving semantics. Although it is true that many entities are already using this type of format, there is still a need to solve the heterogeneous nature of the data, the quality, and the lack of links between different data sets. For this, LODs offer the interconnection of different datasets using standard formats. However, as mentioned above according to Avila-Garzon [4], there is no single and recognized methodology to manipulate LOD, so it is suggested that more research is needed to define a standardized methodology to manage them. Thus, it is necessary to make a systematic review of the technology used for the generation of LOD to serve as a basis for proposing a new framework that public institutions may adopt to satisfy their need to undertake the implementation of LOD processes.

The implementation of the type of technologies mentioned above is necessary for increasing transparency and citizen participation through providing access to LD. Additionally, the research presented in this paper seeks to show the problems that exist during the implementation of LOD, in order to recognize them and, propose technical solutions.

Regarding the research questions, we consider that the applications, tools or technologies to generate and publish LOD are better understood if we identify and expose them as part of a complete and defined process. Therefore, the aim of this research is focused on answering the questions shown in Table 1.

Table 1 Research questions (RQ).

RQ1	RQ2
What are the problems encountered in LOD publication processes?	What are the technical tools applied in LOD publication processes?

For the elaboration of a review protocol, we consider that the most outstanding papers in science are published in high impact scientific databases, such as Scopus and Web of Science (WoS), the latter being the only international and multidisciplinary tool for access such literature a few years ago. However, Scopus, a database founded by Elsevier S.L. in 2004, is gaining strength due to the advantages of navigation and inclusion of 100% of what is indexed in the MEDLINE and EMBASE databases which facilitates fast and open access to information [9]. For this reason, a systematic review is performed in the Scopus database, searching for the keywords: "Linked Open Data", "Linked Open Government Data" and "framework" both in the title and abstract of the publications and grouped in a search string with the Boolean operators AND and OR.

For the systematic review, it is necessary to select the most relevant documents. Thus, we defined the selection criteria shown in Table 2.

Table 2 Selection criteria.

Inclusion Criteria	Exclusion Criteria
Articles and papers published in conferences, in English or Spanish, published between 2018 and 2021.	Documents that are not relevant to answer the research questions, duplicate documents, or documents to which there is no access.

2.2. Conducting the Review

In order to find the technologies used to publish LOD, we first searched for documents containing keywords such as "Linked Open Data" together with another group of words referring to methods, technologies, methodologies, applications, or problems. After reading and refining the search, we decided to simplify the keywords and used the word "framework" which, according to the previous readings, it is observed that the documents that propose a framework for LOD offer a complete approach and detail the applications and technologies used in each stage of the LOD publication process. The final search string is shown in Table 3.

Table 3 Search string.

Search String Text
TITLE-ABS (("linked open data" or "linked open government data") and (framework))

Table 4 shows the number of documents resulting after applying the search criteria. The criteria for selecting papers was based on choosing those studies proposing a specific framework for publishing LOD and detailing technical tools used at each stage.

Table 4 Result after applying the selection criteria.

Inclusion/ Exclusion	Criteria	Documents
Initial search		508
Limited to	Articles and conference papers.	427
Limited to	Language English or Spanish.	415
Limited to	Year of publication > 2017 and < 2022.	99
Excluded	Documents not relevant, duplicated, no access.	5

In the analysis for the extraction of data, we highlighted both, the determination of problems encountered by the author and the technical tools used for the publication of LOD. Table 5 presents the result of the data analysis. To group the technical tools, we adopted the guidelines suggested by Villazón-Terrazas et al. [10] since, according to Saquicela et al. [11], they propose perfectly defined faces in the publication process of Government Linked Data such as Specification, Modeling, Generation, Publication and Exploitation of the data with the objective of becoming an orderly guide for the continuous publication of LD. Table 5 also shows how we related the data extraction criteria to the research questions.

Table 5 Data extraction criteria.

Item	Criteria	Question
Problems	Problems encountered during the LOD publication process are noted or inferred.	RQ1
Data source specification	Name of technical tools used in the Specification stage.	RQ2
Data Modeling	Name of technical tools used in the Modeling stage.	RQ2
Data Generation	Name of technical tools used in the Generation stage.	RQ2
Data Publishing	Name of technical tools used in the Publication stage.	RQ2
Data Exploitation	Name of technical tools used in the Exploitation stage.	RQ2

We only selected technical tools that serve as support in automatic or semiautomatic procedures and that fall within the activities described. We excluded technologies that are only mentioned as examples or do not fall within the described activities. Likewise, configurations and developments that have been made for specific purposes were also excluded.

2.3. Review Report

As the last stage, the results of the review are presented in the Results section of this paper, believing it is convenient to distribute this study in journals and publications oriented to professionals interested in Open Data, Semantic Web, and LOD.

3. Results

This section summarizes the results obtained after the data analysis. The final objective is achieved by answering the formulated research questions. To provide a detailed representation and allow a better understanding, we organize the findings in tables.

3.1. RQ1: What Are the Problems Encountered in LOD Publication Process?

The annotated or inferred problems found in the selected documents are detailed in Table 6.

Table 6 Problems encountered.

Type	Problems Identified	Source
Data Attributes	Heterogeneous data sources.	[12,13,14]
	Data without structured format. Limited access to information. Lack of consistency between data from different sources of origin.	[12]
Tools	JavaScript-based options contain complex technical requirements such as installation and configuration.	[12]
	Lack of integrated solutions to help non-expert users to publish LOD. Lack of integrated solutions for process tracking and semantic visualization.	[14]
	Lack of formal methodologies to support the LOD implementation process. Lack of platforms to support all phases of LOD publication.	[6]
	Problem of comprehensiveness of current tools in the LOD publication process.	[11]
	Automatic procedures to extract information can be difficult.	[12]
Human Capacity	Difficulty of information manipulation for non-expert users.	[13]
	A high level of conceptual and technical knowledge is needed for the correct generation and publication of LODs.	[6]

In this section, we present the results derived from the analysis of the selected papers: [6,11,12,13,14], in which the authors have pointed out the challenges associated with LOD publication and the need to propose a new framework to facilitate this process. The findings were organized into three main categories: problems related to data attributes, difficulties in the tools used, and obstacles linked to human capacity. The objective is to provide a comprehensive and informed view of the challenges facing LOD implementation. By effectively addressing these problems, it will promote the publication of LOD in various fields, thus boosting the development of the Semantic Web as a whole.

Our results reveal that one of the key problems within the "Data Attributes" category in the LOD publishing process is the heterogeneity of the data, which necessitates the use of tools to preprocess the data. Another problem we highlighted within the "Tools" category is the absence of integrated solutions to help users publish LOD, which led us to conduct this systematic review to get an overview of the steps needed to publish LOD and explore possible tools to support the process. Within the "Human Capacity" category, we highlighted the difficulty that less experienced users have in publishing LODs, which may affect the quality and quantity of published LODs.

3.2. RQ2: What Are the Technical Tools Used in LOD Publication Process?

To answer this question, we conducted a thorough analysis of each document and then noted which tools were used in each document to support the LOD publication process. Table 7 summarizes the tools identified as used in each phase of the LOD publication process.

Table 7 Technology used in the publication of LOD.

Phase	Activity	Technology	Source
Specification	Data mapping and preprocessing sources	OpenRefine	[6]
		Pentaho Data Integration (Kettle)	[11,12]
		Web scratching systems	[13]
		ETL Tools	[14]
Modeling	Vocabulary mapping	BIBO, DCTERMS, FOAF, RDAA	[11]
		RDF Data Cube Vocabulary	[12]
		Ethnic groups in Thailand ontology, Dublin Core Metadata, Web Ontology Language, cross-domain DBpedia Ontology	[13]
		Water Supply Network Ontology	[14]
	Ontology editor tool	Hozo - Ontology Editor	[13]
		Protégé	[14]
Generation	Cleaning and data transformation	OpenRefine	[6,12,13]
		Pentaho Data Integration (Kettle)	[11]
		Apache Jena	[13,14]
	Repositories used as external source data	Freebase, NCBI taxonomy	[6]
		DBpedia	[6,13,14]
		Dspace	[11]
		Wikidata	[12,13,14]
		Geonames	[12,14]
		VIAF	[14]
	Linking to other data sources	Silk Link Discovery Engine	[6]
		Pentaho / Silk plugin	[11]
		OpenRefine	[12,13]
		Mix'n'match	[14]
	Validation	RDF NTriples/Turtle Validator	[12]
		Stardog's Integrity Constraint Validation, W3C RDF validator	[14]
Publication	Data storing and publication	Pubby server	[6]
		Apache Jena Fuseki	[6,13]
		Pentaho / Fuseki Loader plugin	[11]
		RDF4J server, DataHub	[12]
		Jetty and Pubby	[13]
		Stardog	[14]
Exploitation	Consumption and data visualizing	Neo4j datastore	[6]
		Pentaho / ELDA Loader plugin	[11]
		CubeViz.js	[12]
		SPARQL query	[6,11,12,13,14]

The table shows the phases of the LOD publication process, the activities carried out in each phase, the technical tools used, and their respective sources. Next, we explain the LOD publication process with details on the tools that we have found in the analysis of the selected articles and that have been used in each phase, answering research question 2 (RQ2). The purpose of presenting these results is to provide a comprehensive perspective that explains the full cycle of LOD publication and the tools that various authors have chosen to develop their frameworks. This will provide valuable input to create a new and updated framework that suits the needs of a government organization and, above all, overcomes the challenges encountered in LOD management.

In the Specification phase, data mapping and preprocessing of sources are performed to ensure the quality and relevance of the data to be published. Extraction, Transformation, and Loading (ETL) tools, such as OpenRefine (formerly Google Refine) and Pentaho Data Integration (Kettle), are used. OpenRefine allows working with messy data, cleaning, transforming, and enriching it with external data [15]. On the other hand, Pentaho offers ETL capabilities for capturing, cleansing and storing data in a uniform, consistent, accessible, and relevant format for end users, including Internet of Things (IoT) technologies [16].

In the Modeling phase, ontologies are created by reusing existing ones or starting from scratch. The tools mentioned include Protégé, a Stanford University project, which helps to build reusable ontologies and knowledge-based systems [17], and the Hozo tool, with ontology editing functions [18].

The Generation phase involves data cleansing and transformation activities, for which tools such as OpenRefine, Pentaho Data Integration (Kettle), and Apache Jena are used. Apache Jena provides information as a collection of RDF triples that are contained in a data structure [19]. It then links to structured data repositories like Wikidata, Dbpedia, and Geonames to create a linked and enriched information network that improves data interoperability and accessibility. To search for links to other sources, tools like Mix'n'match, allowing matching entries from external databases with Wikidata elements [20], are used. Silk, an open-source framework for integrating heterogeneous data sources using RDF links [21] is also applied. To ensure the quality, consistency, and usability of the linked data the validation activity is performed, in this section, we mention the RDF tools NTriples/ Turtle Validator, which verifies Turtle and Ntriples documents on syntax errors and XSD data types through the command line [22]; as well as Stardog's Integrity Constraint Validation, a function to enforce data integrity along with knowledge graph correctness and consistency [23] and W3C RDF Validator, an RDF validation service based on Another RDF Parser (ARP) [24].

In the Publication phase, some key activities are carried out such as data storage and publication to facilitate access to the LOD and its subsequent exploitation by the community, in these activities we found tools like RDF4J, a Java framework for RDF data processing and management [25]. Datahub, a platform for publishing, deploying, and sharing data [26], and Apache Jena Fuseki, chosen to manage the RDF data store. This tool operates as a SPARQL server, serving a dual purpose: it provides a space for storing information in RDF format and, at the same time, establishes an access point for conducting specific queries about this data [6]. To facilitate data availability through HTTP, a front-end of LOD is employed and configured. Specifically, the implementation of Jetty and Pubby, components that are part of the LOD API specification, is chosen. This frontend ensures data accessibility via HTTP and enables content mediation, thereby allowing users to request information in a variety of formats [13]. Pubby server aims to facilitate internet access to RDF triples. Pubby provides a Linked Data interface for both local and remote SPARQL protocol servers [6]. Additionally, Stardog, a graphical platform with data infrastructure that can be virtualized and accessed by applications, and Business Intelligence tools [27].

Exploitation is the final phase of the LOD publishing process, where tasks such as data consumption through federated queries and visualization of the obtained data are developed with the intention of generating knowledge, facilitating decision making, or creating intelligent applications. CubeViz.js, which allows the exploration and visualization of statistical data in RDF, uses the RDF DataCube vocabulary and is written entirely in Javascript [28]. SPARQL is also used to express queries over various data sources stored in RDF [29]. In addition, NEO4J is a data graph platform that enables the creation of intelligent applications and machine learning workflows [30]. Finally, we take note of the utilization of Elda, an implementation in Java of the LD API that provides access to RDF data through RESTful URLs, which are translated into queries to a SPARQL endpoint [31].

3.3. Analysis of the Predominant Tools in LOD Publication Process

We analyzed the technical tools that we consider predominant in Table 7, namely Pentaho Data Integration (Kettle) and OpenRefine. From the papers reviewed, we can infer that the mentioned characteristics guided the authors in the selection of the tool used in each framework. First, in Table 8, we present the characteristics detailed by the authors of the papers listed in Table 7. Then, in the Discussion section, we will provide a comparison, highlighting the similarities between these tools and detailing features relevant to our study.

Table 8 Pentaho and Open Refine Features.

Type	Pentaho Data Integration (Kettle)	Open Refine
Data Preparation	Cleans and normalizes data from heterogeneous data sources. Allows access, preparation, combination, and analysis of unstructured data. Supports different data formats (e.g., csv, excel, relational databases, web services, etc.).	It is a tool in terms of working with heterogeneous data, transforming it into a uniform vocabulary, and enriching it with external repositories. Unifies and cleans data by correcting errors, removing duplicates, and preparing it for transformation. It is an open-source tool that offers data formatting and sorting capabilities.
Data Transformation	Data is transformed to a standard format for data exchange within LD, i.e., RDF. It provides a set of data cleansing plugins designed to support data transformation and manipulation.	It can transform source data into RDF data cube vocabulary. Allows transforming CSV files into RDF according to standard vocabularies. Transforms raw data into a machine-readable format. It performs a graphical mapping from the project to an RDF skeleton.
Data Interconnection	It maps the extracted and pre-processed data from the data sources to the selected ontology vocabulary.	Adds to the data structure the provision of interconnection with other data sources.
Flexibility and Scalability	Allows extending the tool to support other input data formats by implementing ad-hoc plugins. Native plugins support loading data from different sources and deliver them as data tables, which is the data structure used in Kettle.	It can be complemented with extensions, such as the RDF extension, to provide additional functionalities.
User Interface	Enables the delivery of a comprehensive solution to support the entire LOD life cycle.	Offers an easy-to-use user interface, especially for non-technical and non-expert users. It is an open-source desktop application. Allows data to be grouped, which helps to better understand the data.

Table 8 summarizes the characteristics Pentaho Data Integration (Kettle) and OpenRefine, extracted from the analyzed documents in which such tools were used. For a better understanding, they have been categorized according to different aspects, such as data preparation, transformation, interconnection, flexibility, scalability, and user interface.

We can observe that both tools share similar characteristics, which are useful in data cleaning and transformation processes. Additionally, other features also exhibit some concordance. However, each tool has its particularities that necessitate further analysis in future studies

Our main objective in this analysis was to highlight the functionalities of both tools. In the following section, we will delve into a detailed comparison and discussion of these features, with the purpose of providing a comprehensive and objective view to enable informed decisions when selecting tools for LOD implementation.

4. Discussion

4.1. Problems Encountered in LOD Publication Process

Table 6 provides an overview of the problems identified during the implementation of the LOD publication life cycle, some of which have common particularities that led us to group them by the type of problem for a better understanding.

In the "Data Attributes" category, it is noted that the heterogeneity of data sources and the lack of structure and consistency between data represent obstacles in the data preparation. This activity is critical, as the quality and interoperability of the LOD depend on the standardization and homogenization of the data. To overcome these challenges, it is necessary to develop methodologies and standards that facilitate the transformation and standardization of the data in a way that conforms to the LOD principles.

Regarding the "Tools" category, challenges are identified that affect the entire LOD publication process. JavaScript-based tools are presented as a complex option, which could hinder adoption by non-expert users. This highlights the need to develop accessible and user-friendly solutions to facilitate publishing LOD. Such need highlights two issues. On the one hand, the lack of integrated solutions, methodologies, and platforms covering all stages of the LOD lifecycle may result in a fragmented, unfinished, and less efficient implementation. On the other hand, the problems of understanding the current tools lead to the necessity of developing easy automatic procedures for information extraction. Consequently, it is crucial to encourage the design of frameworks that guide users to determine the precise tools and correct steps to ensure the correct publication of LOD even for non-expert users.

Finally, the "Human Capacity" category highlights the challenges in users' knowledge and expertise in handling technical tools that support the LOD publication process. The manipulation of technical tools is shown as a difficulty for non-expert users, highlighting the need for more intuitive interfaces that facilitate the publication of LOD. In addition, the requirement of a high level of conceptual and technical knowledge to generate and publish LOD may limit the widespread adoption of this technology. In this sense, in addition to investing in training programs, it is needed to propose, develop, or improve existing tools to facilitate their use.

4.2. Technical Tools Used in LOD Publication Process

Although we previously justified the adoption of the guidelines proposed by Villazón-Terrazas et al. [10] in the methodological section, it is important to highlight that there are several contributions related to frameworks and their methodology proposed for the publication of LOD. For example, the author Saquicela et al. [11], mentions some of them, such as Linked Data Integration Framework (LDIF), Information Workbench (DataOps), and Optique. These contributions lead us to the need to further extend our studies

Table 7 provides a life cycle perspective on LOD publishing. The presentation of technical tools grouped by each phase provides a structured view of the technological landscape in LOD publishing, allowing researchers and practitioners to understand the interaction of the tools throughout the process.

Aspects such as data quality and relevance should be considered in LOD publishing. The technical tools identified in this study offer features that support such aspects. For example, OpenRefine and Pentaho Data Integration (Kettle) were commonly used in the Specification phase, addressing data mapping and preprocessing activities.

In the Modeling phase, we observed a trend toward the creation of ontologies for LOD publication. In the analysis of the selected papers, it is evident that existing ontologies were used instead of starting from scratch. Protégé and Hozo are tools used to support the creation and editing of ontologies. However, due to the limitations of this study, we cannot conclude a clear trend in this phase, suggesting that further research on this topic is needed.

In the Generation phase, data cleaning and transformation activities are carried out. Frequent use of OpenRefine and Pentaho for these tasks is again observed, indicating their dominance within the LOD publishing process. The results also show that several external data repositories, such as Wikidata, Geonames, and DBpedia, were widely used, denoting that combining these repositories can enrich the linked information network and improve data interoperability in the LOD environment. Additionally, the use of Silk as part of a framework in some papers reinforces the idea of using this tool to strengthen the integration of data from various sources. However, the tools presented in Table 7 and used for validation do not demonstrate a clear trend in this study, suggesting that further research is needed to reach a stronger conclusion.

To facilitate access to LODs, data storage, and publication activities are carried out for subsequent exploitation. In the Publication phase, the Apache Jena Fuseki tool prevails. This tool functions as a SPARQL server, ensuring that the published linked data is accessible through SPARQL queries for any user. We have also identified another common tool in some documents: Pubby Server. This tool enables users to interact with SPARQL endpoints following LD principles. This justifies the common use of SPARQL queries in the Exploitation phase. Therefore, according to our study, Apache Jena Fuseki, Pubby Server, and SPARQL queries used together are fundamental to the LOD publishing process.

4.3. Predominant Tools in LOD Publication Process

After completing the analysis of technical tools used in the LOD publication process, it can be concluded that two ETL tools predominate: Pentaho Data Integration (Kettle) and OpenRefine. However, other tools have not been mentioned in the documents analyzed, but which are experiencing a growth in use in recent years. These include Apache Hop, as an alternative to Pentaho, and Apache Spark, used for distributed data processing. In addition, other cloud-based ETL tools and services, such as AWS Glue managed by Amazon Web Services (AWS), and Data Factory managed by Microsoft Azure, have been observed and are also gaining popularity. In an extension of this research, it would be interesting to explore such additional tools and examine how they equate with those mentioned above.

In Table 8, we can observe the features of Pentaho and OpenRefine in the context of their relevance to the LOD publishing process. However, it is essential to keep in mind that this table only presents the features mentioned by the authors, there are functionalities that can be evidenced through the use of the tools in different case studies. These additional functionalities could be investigated in future work. Therefore, in this discussion, we focus on the characteristics found in the analysis of the selected papers.

First, both tools show features for working with, cleaning, normalizing, and unifying heterogeneous data. As shown in Table 7, Pentaho and OpenRefine offer similar functionalities in this regard. Both tools allow users to access, prepare, and analyze unstructured data, which can be beneficial in managing complex data sets. Furthermore, both Pentaho and OpenRefine support various data formats, which would allow users to work with diverse data sets.

In addition, both Pentaho and OpenRefine share the common goal of transforming data to a standard format, namely RDF, which is a necessary step for data publishing and subsequent linking. We can also note that both tools allow data interconnection and integration with other data sources.

Both Pentaho and OpenRefine allow the extension of the tool by implementing ad-hoc plugins and extensions respectively. This could provide users with the ability to add new functionality and tailor the tool to their specific requirements.

A notable feature of Pentaho is that it is an open-source desktop application, while the information provided does not specify the platform type for OpenRefine. This underscores the need for more specific studies to thoroughly analyze the tools.

Finally, we highlight that Pentaho offers an end-to-end solution to support the entire LOD lifecycle. This allows users to work on a single platform for multiple tasks related to LOD publishing. Moreover, a notable feature of OpenRefine is its user-friendly interface, specifically targeted at non-technical and non-expert users. This would allow more users to use the tool without the need for extensive technical knowledge.

5. Conclusions

In this paper, we presented a systematic literature review on problems encountered in LOD publishing and the tools used in each phase of the LOD publishing life cycle. After performing the search on Scopus and applying the inclusion and exclusion criteria, we analyzed 99 documents within which we found five publications that specified a complete LOD publication life cycle and the technical tools used in each phase. Based on the selected papers, we identified the problems encountered in LOD publication, among which the following prevail: heterogeneous data sources; the functionality of technical tools, which result complex for the use by non-expert users; and the lack of comprehensive solutions and methodologies to support the LOD publication process.

Regarding the technical tools that support the LOD publication process, we have made a synthesis and observed that in the Specification and Generation phases, ETL tools such as Pentaho Data Integration (Kettle) and OpenRefine stand out. Among the most used repositories as external data sources are Wikidata, Geonames, and DBpedia. For link discovery, we observed that Silk has a prevalence in several documents. In the Publishing phase, the most utilized applications are Apache Jena Fuseki and the Pubby server. For the Exploitation phase, we identified the SPARQL query tool as the most commonly employed.

In this work, we have limited ourselves to using a single LOD publication methodology as a guide to represent the methodological tools in their different phases. However, it is important to note that there are other proposed methodologies and frameworks. To overcome these limitations, the possibility of investigating the combination of different frameworks and methodologies could be considered, taking advantage of the strengths of each to build a more comprehensive solution.

Among the lifecycles proposed by various authors, we emphasize the predominant use of Pentaho Data Integration (Kettle) and OpenRefine as tools. One advantage of Pentaho is its comprehensive solution, supporting the entire LOD lifecycle. On the other hand, OpenRefine stands out for its user-friendly interface, specifically designed for non-technical and non-expert users

The use of the Scopus database proved to be sufficient to achieve the objectives and cover the scope of the present article. However, to expand the research with new objectives, we propose as future work, to explore other open spaces containing articles and scientific journals, including those that do not require payment of article processing charges (APCs), and that present a vast number of publications, as in the case of DOAJ. In future research, it is also recommended to additionally consider other languages and the inclusion of articles and journals from databases such as Dimensions or Google Scholar, as well as to complement the search for information with other reliable sources, such as IEEE, SPRINGER as well as specialized journals such as Semantic Web Journal. These additional sources will provide access to knowledge and relevant scientific findings in the area of study, thus strengthening the evidence base used in the research. Considering the growing prominence of new technical tools in recent years, it is essential to carry out, as future work, a comparative analysis of tools such as Apache Hop, Apache Spark, AWS Glue, and Data Factory, in contrast to those mentioned in this research: Pentaho Data Integration (Kettle) and OpenRefine.

Finally, we can conclude that one of the challenges we face is the lack of integrated solutions and methodologies to support LOD publication processes, especially for non-expert users. For this reason, our future work will be to conduct research aiming at proposing a framework to standardize LOD publication processes in public institutions to be used by both, expert and non-expert users.

Competing interests

The authors have declared that no competing interests exist.

Authors' contribution

Jairo Silva Aguilar conceived the idea, conducted the systematic literature review, analyzed the results, and wrote the manuscript; Romel Torres and Elsa Estevez provided methodological advice, analyzed the results, revised and corrected the manuscript. All authors read and approved the final manuscript.

Acknowledgements

This research has been supported by the Government of Ecuador through the Secretaría de Educación Superior, Ciencia, Tecnología e Innovación.

References

[1] E. Ruijer, F. Détienne, M. Baker, J. Groff, and A. J. Meijer, "The Politics of Open Government Data: Understanding Organizational Responses to Pressure for More Transparency," *Am Rev Public Adm*, vol. 50, no. 3, pp. 260–274, 2020, doi: 10.1177/0275074019888065.

[2] E. Ruvalcaba-Gómez, "Datos abiertos como política pública dentro del Gobierno abierto," *Revista sobre el Estado, la administración y las políticas públicas*, vol. 3, no. 2, pp. 99–116, 2019.

[3] T. Janowski, E. Estevez, and R. Baguma, "Platform governance for sustainable development: Reshaping citizen-administration relationships in the digital age," *Gov Inf Q*, vol. 35, no. 4, pp. S1–S16, 2018, doi: 10.1016/j.giq.2018.09.002.

[4] C. Avila-Garzon, "Applications, methodologies, and technologies for linked open data: A systematic literature review," *Int J Semant Web Inf Syst*, vol. 16, no. 3, pp. 53–68, 2020, doi: 10.4018/IJSWIS.2020070104.

[5] T. Berners-Lee, "Linked Data," 2006. `https://www.w3.org/DesignIssues/LinkedData.html` (accessed Aug. 05, 2021).

[6] R. Abida, E. Hachicha Belghith, and A. Cleve, "An End-to-End Framework for Integrating and Publishing Linked Open Government Data," *Proceedings of the 29th IEEE International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE 2020). IEEE Computer Society Press.*, 2020.

[7] M. A. Espinosa C, E. Romero R, L. Y. Flórez G., and C. D. Guerrero, "DANDELION : Propuesta metodológica para recopilación y análisis de información de artículos científicos . Un enfoque desde la bibliometría y la revisión sistemática de la literatura .," *Iberian Journal of Information Systems and Technologies*, pp. 110–122, 2020.

[8] B. Kitchenham and S. Charters, "Guidelines for performing Systematic Literature Reviews in Software Engineering," *Keele University and Durham University Joint Report*, no. Ver. 2.3, EBSE-2007-01, 2007.

[9] C. A. Calvache-Mora and M. A. Ríos-Ramírez, "Bibliometric analysis of the scientific production found in Scopus and Web Of Science about physiological vocal rehabilitation," *Revista de Logopedia, Foniatria y Audiologia*, vol. 38, no. 3, pp. 120–129, 2018, doi: 10.1016/j.rlfa.2018.04.004.

[10] B. Villazón-Terrazas, L. M. Vilches-Blázquez, O. Corcho, and A. Gómez-Pérez, "Methodological Guidelines for Publishing Government Linked Data," in *D. Wood (ed.) Linking Government Data*, 2011, pp. 27–49. doi: 10.1007/978-1-4614-1767-5_2.

[11] V. Saquicela *et al.*, "LOD-GF: An Integral Linked Open Data Generation Framework," *Advances in Intelligent Systems and Computing*, vol. 884, pp. 283–300, 2019, doi: 10.1007/978-3-030-02828-2_21.

[12] P. Escobar, G. Candela, J. Trujillo, M. Marco-Such, and J. Peral, "Adding value to Linked Open Data using a multidimensional model approach based on the RDF Data Cube vocabulary," *Comput Stand Interfaces*, vol. 68, no. February 2019, p. 103378, 2020, doi: 10.1016/j.csi.2019.103378.

[13] W. Chansanam, K. Tuamsuk, J. Chaikhambung, and S. Sugimoto, "Linked open data framework for ethnic groups in Thailand learning," *International Journal of Emerging Technologies in Learning*, vol. 15, no. 10, pp. 140–156, 2020, doi: 10.3991/ijet.v15i10.13337.

[14] P. Escobar, M. del M. Roldán-García, J. Peral, G. Candela, and J. García-Nieto, "An ontology-based framework for publishing and exploiting linked open

data: A use case on water resources management," *Applied Sciences (Switzerland)*, vol. 10, 2020, doi: 10.3390/app10030779.

[15] OpenRefine, "OpenRefine," 2021. `https://openrefine.org/` (accessed Apr. 05, 2021).

[16] Pentaho, "Pentaho Data Integration," 2020. `https://help.pentaho.com/Documentation/9.1/Products/Pentaho_Data_Integration` (accessed Apr. 05, 2021).

[17] M. A. Musen and the Protégé Team, "The Protégé Project: A Look Back and a Look Forward," *AI Matters*, vol. 1, no. 4, pp. 4–12, 2015, doi: 10.1145/2757001.2757003.

[18] Kouji KOZAKI, "Hozo - Ontology Editor," 2020. `http://www.hozo.jp/` (accessed Aug. 01, 2021).

[19] Apache Jena, "The core RDF API," 2021. `https://jena.apache.org/documentation/rdf/index.html` (accessed Apr. 05, 2021).

[20] Mix'n'match, "Mix'n'match." `https://mix-n-match.toolforge.org/` (accessed Apr. 05, 2021).

[21] SilkFramework, "Silk - The Linked Data Integration Framework." `http://silkframework.org/` (accessed Apr. 05, 2021).

[22] IDLabResearch, "Turtle Validator," 2020. `https://github.com/IDLabResearch/TurtleValidator` (accessed Jul. 30, 2021).

[23] Stardog, "Data Quality Constraints," 2021. `https://docs.stardog.com/data-quality-constraints` (accessed Jun. 01, 2021).

[24] World Wide Web Consortium, "RDF Validation Service," 2007. `https://www.w3.org/RDF/Validator/documentation` (accessed Jul. 30, 2021).

[25] Eclipse Foundation, "Eclipse RDF4J," 2021. `https://rdf4j.org/` (accessed Jun. 01, 2021).

[26] DataHub, "DataHub Open Data," 2018. `https://datahub.io/` (accessed Jun. 03, 2021).

[27] Stardog, "Stardog," 2021. `https://www.stardog.com/` (accessed Jun. 01, 2021).

[28] AKSW, "cubevizjs," 2021. `https://github.com/AKSW/cubevizjs` (accessed Jun. 08, 2021).

[29] World Wide Web Consortium, "SPARQL Query Language for RDF," 2008. `https://www.w3.org/TR/rdf-sparql-query/` (accessed Apr. 05, 2021).

[30] Neo4j, "neo4j," 2021. `https://neo4j.com/` (accessed Jun. 04, 2021).

[31] Epimorphics, "ELDA," 2021. `https://github.com/epimorphics/elda` (accessed Jun. 04, 2021).

Citation: J. H. Silva-Aguilar, R. Torres T. and E. Estevez. *Publication of Linked Open Data – A Systematic Literature Review for Identifying Problems and Technical Tools Supporting the Process.* Journal of Computer Science & Technology, vol. 23, no. 2, pp. 179-189, 2023.
DOI: 10.24215/16666038.23.e16
Received: April 16, 2023 **Accepted:** August 14, 2023.