A Survey of Current Approaches for Transforming Open Data to Linked Data

Amina Meherhera, Imane Mekideche, Leila Zemmouchi-Ghomari, Abdesamed Réda Ghomari

A Survey of Current Approaches for Transforming Open Data to Linked Data

MEHERHERA Amina[1], MEKIDECHE Imane[1], Dr. ZEMMOUCHI GHOMARI Leila[2], and Pr. GHOMARI Abdesamed Réda[1]

[1] LMCS - École Nationale Supérieure d'Informatique (ESI ex.INI), Oued Smar, Algiers.
[2] École Nationale Supérieure de Technologie (ENST), Bordj El Kiffan, Algiers.
fa_meherhera@esi.dz , fi_mekideche@esi.dz , leila.ghomari@enst.dz , a_ghomari@esi.dz

Abstract. There are billions of open datasets published on the web. Nevertheless, most of these data are merely available in a human-readable format and are not interlinked. Thus, the full potential of the web is not being used.
With the aim to address this issue, Tim Berners Lee proposed some of the best practices to publish data on the web, known as the linked data principles, enabling machines to interpret the semantics of data correctly and therefore infer new facts and return relevant results of web research. This paper is a systematic literature review on existing approaches to transform web open data to linked data. A classification of these approaches is proposed, as well as a synthesis study to highlight approaches' main trends and challenges.

Keywords: Linked Data · Open Data · Transforming Approaches.

1 Introduction

The world wide web counts today billions of web pages and documents that are full of data; the classic web shows significant limitations in terms of organising, extracting, and processing the available data on documents. Thus, the necessity of moving to another version of the web, the semantic web, introduced by Tim-Berners-Lee in 2009, which enables data providers to publish their data efficiently so that it can be machine-readable and facilitate its reuse.
One of these techniques is the linked data paradigm, which aims to enable uniform practices for publishing interconnected data on the web using semantic web technologies [11].

This transforms the classical web from a network of documents to a network of data allowing software agents and machines to interpret the data.

The adoption of the interconnected data principles by a large number of companies, research centers, and institutions around the world has led to the creation of a global data space with interconnected data from different domains, such as people, companies, books, films, statistical and scientific data. [11].

This data network is called the "Linked Open Data (LOD) Cloud", a massive network of published datasets available via the existing web infrastructure and interconnected according to the linked data principles.
The linked data principles provide general guidance for making a given dataset available on the web in linked data format. However, there are several different tools, methods, and techniques for generating and publishing these datasets.
In this paper, we present a bibliographical study of linked data in its context, the semantic web, as well as the set of existing approaches and current efforts to perform the transformation from open data to linked data.
For this aim, we conduct a systematic survey on the literature related to transforming web open data to linked data. Our aim is to provide new interpretations of existing literature.
Our survey is composed of four sections. Section 2 covers basic concepts of the paper domain. Section 3, describes the process of systematic survey. Section 4 presents the selection and the study of existing approaches and initiatives to transform open web data to linked data. We classified them according to some criteria, such as automation degree, supported files format and application domain. We assessed the methodological quality of studies and discussed the different initiatives, where we will identify the significant limitations and challenges they face. Section 5 concludes the paper and presents its perspectives.

2 Background

In this chapter, we introduce the basic concepts of the semantic web and linked data.

2.1 Definitions:

The semantic web is an extension of the World Wide Web that aims to publish data in a format that is machine-readable and thus allows computers to intelligently search, combine and process web content based on its semantics. [12]
It is based on a set of best practices for publishing and interlinking structured data of multiple sources on the web. Theses best practices are known as the linked data principles mentioned in section 2.2.[4]
Linked data makes the World Wide Web into a global database that we call the web of data. Developers can query linked data from multiple sources at once and combine it on the fly, something challenging to do with traditional data management technologies. [28]
We can't talk about linked data without Open data which can be summed up in the statement: "Open means anyone can freely access, use, modify, and share for any purpose " [3].
Publishing Data on the web as linked data requires some special technologies: RDF for representing data and SPARQL for querying data are the most important semantic web tehnologies.

[3] `http://opendefinition.org/`

RDF [4] is a language for representing information about resources in web. It is based on the idea of identifying things using web identifiers (called Uniform Resource Identifiers, or URIs) and describing resources in terms of simple properties and property values. This enables RDF to represent simple statements about resources as a graph of nodes and arcs representing the resources, their properties and values. [5] In RDF, the description of a resource is represented as a number of triples. The three parts of each triple are called: **subject** (the URI identifying the resource), **object** (can be either a literal value or a URI) and **predicate** (indicates the relation between the subject and the object).
SPARQL[6], is a set of specifications that provide language and protocol to query and manipulate RDF graph content on the web or in an RDF store[7]

2.2 Linked data principles

Tim-Berners-Lee, the inventor of the web and the initiator of the Linked Open Data project, introduces the four linked data principles:

1. Use URIs as names for things.
2. Use HTTP URIs, so that people can look up those names.
3. When someone looks up a URI, provide useful information, using the standards (RDF, SPARQL).
4. Include links to other URIs, so that they can discover more things.

2.3 Linked Data applications

"Linked data successes are difficult to see because they are under the hood" [28]. Linked data has been widely used since its inception, we cite here some of the success stories of linked data: [29]

- **Google Knowledge Graph**: it was introduced by Google in 2012 to enhance search results and allow users to extend their knowledge by searching for related topics. [21]. The google snippet that is shown in the right side when searching for something on Google is one of its use cases.
- **Open Government Data (OGD):** a collaboration between US, UK, France and Singapoure governments to publish machine readable datasets. Actually, the data.gov offers more than 5 billions RDF triples.
- **BBC website:** One of the most known success stories of linked data is the BBC website that uses the linked data principles to publish their data.

[4] (Resource Description Framework)
[5] `https://www.w3.org/TR/2014/NOTE-rdf11-primer-20140624/` , Feb 2020
[6] SPARQL Protocol And RDF Query Language
[7] `https://www.w3.org/TR/2013/REC-sparql11-overview-20130321/`, Feb,2020

3 Methodology

In our work, we followed a systematic literature approach: a synthesis of the scientific literature in response to a specific question. It uses explicit methods for **restrictively** searching, selecting, and analyzing data. Our approach is based on the guidelines proposed by [15] and [7]. The procedure is:

1. Definition of the terms/questions of the research.
2. Selection of the sources (digital libraries) to be searched.
3. Research and selection of studies.
4. Data extraction and analysis.
5. Evaluation of the methodological quality of the studies.
6. Synthesis of results.

3.1 Defining the terms/questions of the research:

The identification of research questions is essentially what distinguishes systematic research from traditional research. The goal of this survey is to analyse existing initiatives, tools and approaches to transform web open data into linked data. We therefore define the following as generic research questions:

- What are the existing approaches to transform open data to linked data?
- How can they be classified ?
- What technical aspects, features and functions are supported in the existing approaches ?
- What are the current challenges and general trends related to the transformation of open web data to linked data?

Keywords in search : Transform to linked data, Convert to linked data, Produce linked data , From open data to linked data.

3.2 Selecting the sources (digital libraries) to be searched.

In order to cover the widest possible range of relevant publications, we have identified and used the most widely used electronic libraries, namely : ACM Digital Library, IEEE Xplore, ScienceDirect, Google Scholar, HAL, SemanticScholar.

4 Literature Review

4.1 Research and selection of studies :

In order to select the most relevant articles, we have proceeded to **gathering**, **analysing** and then **filtering** articles in four steps:

1. Searching for articles using the keywords mentioned in section 3.1.
2. Selecting articles with titles related to our study: the result of this step was: **42 articles**.

3. Quick reading of the abstract, introduction and conclusion to select according to **reliability** (number of citations, variety and quality of references, reputation of the author, source of the article) and **relevance** (The article proposed an approach to transform open data to linked data or contains relevant information for our study): The result of this step was: **22 articles**.
4. Finally, we have read the articles and applied an other filter on them to select those which:
   - proposed a **new** approach for transforming open data to linked data
   - proposed an approach producing linked data of **good quality** (respecting the four linked data principles).
   - performed tests on their approach and good results have been revealed.
   - explained the steps of their approach in a non ambiguous way.

   The result of this step was: **12 articles**.

Figure 1 represents the process that we have followed to filter and select articles.

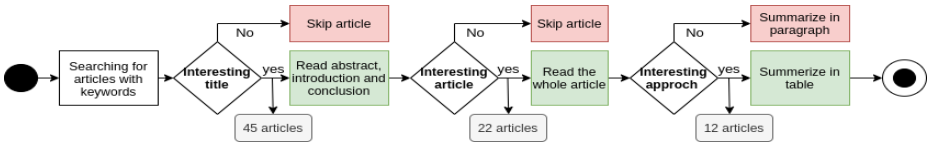

Fig. 1: The followed process of articles selection

4.2 Data extraction and analysis :

In this section, we will talk about the approaches resulting from the three filters:

- [**16**] : proposed an incremental approach of transforming governmental datasets available in data.gov to linked data by converting to RDF format and then using existing vocabularies for interlinking.
- [**13**]: the authors proposed a step-by-step process for modelling data, converting and publishing new datasets that has enabled many governments to expose data for further visualisation, reuse and analysis
- **Linked data life cycles**[8]: the authors state that existing data management methods which assume control of data, schema and data generation cannot be used in the web environment because of its open and decentralised nature. This is an approach suggested by the W3C group as it covers all stages of the linked data life cycle, but unfortunately it focuses on government data.
- [**17**] : proposed a technique to automatically produce good quality linked data by deducing the semantics of columns headers, cell values and relations between columns inferred by knowledge base from different data sources such as the LOD cloud .
- [**9**]: creates a RDF Triple Store populated with triples produced from text documents to extract relations between terrorists or terrorism organisms.

[8] `https://www.slideshare.net/mediasemanticweb/linked-data-life-cycles`

- **[27]** : covers all stages of government linked data life cycle.
- **[2]**: provides software tools for the integration of the very heterogeneous LOD2 data in a coherent framework.
- **[1]**: worked on transforming data coming on streaming of AEMET [9] dataset to linked data by developing a specific ontology for the dataset.
- **[3]**: it aims to integrate city data into a common data model using semantic web technologies in order to help city managers with their decisions by providing automated analytical support.
- **[25]**: an approach based on conceptual mapping between CSV components and ontology components and keeping relations between columns by proposing algorithms to detect the datatype or the object.
- **[23]** : presented a platform that aims to change the paradigm of open government data portals in Romania by publishing data in LOD format providing an easy way for developers to create applications without having to process the initial files (by providing a SPARQL engine).
- **[19]**: the goal of the datalift project was to foster the emergence of the web data by industrialising and facilitating the structuring, publication, interlinking and use of critical mass of data, providing access to spatial datasets from oceanographic archives on the web.
- **[14]** : proposed a framework to link and interrogate data published as csv files on the web (the data.gov plateform) .

After briefly describing the contribution of each article , we classified the approaches according to different criteria:

- **Supported files formats**
- **Application domain**
- **Degree of automation**

The result of the classification are represented in the figure 2

4.3 Evaluation of methodological quality of the studies :

Some approaches analyzed in this article, like [13] and [27] included most steps of the generally accepted linked data life-cycle but missed guidelines on how to use the generated linked data. They do not provide detailed guidelines for publishing the generated linked dataset on the web.
Another point to consider is that most of the studies used explicit data sources and published them on the web as linked data without proper semantic description and documentation. We noticed that most are focused on public data.
After analysing and comparing existing approaches in literature and selected in this article, we present the most common limitations:

- Some approaches are specific to a domain or a dataset and are hard to be adapted to another domain or other data available in the web.

[9] Agencia Estatal de Meteorología

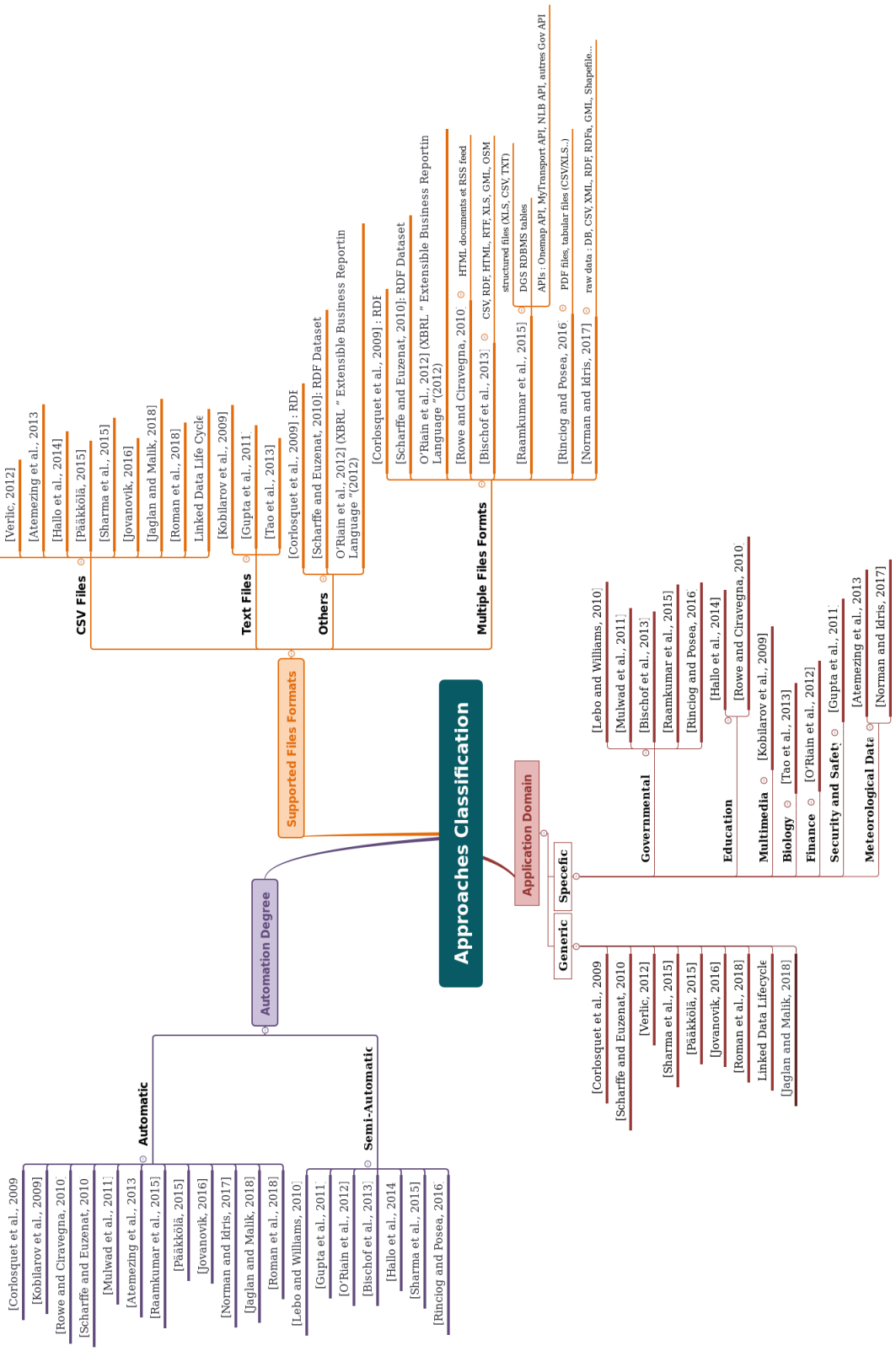

Fig. 2: Classification of selected articles

- Some approaches accept restricted files format as entries (CSV files are the most supported).
- Some approaches are not linking to other datasets and thus not complying with the fourth linked data principle.
- Some approaches have some steps that need human intervention (are semi-automatic).

4.4 Synthesis of the results :

The linked data principles provide general guidance for making a given dataset available on the web in linked data format. However, there are several different tools, methods and techniques for generating and publishing these datasets, and their application depends on the type, the nature of data and a list of other factors. These methods and techniques are brought together in several linked data life-cycle methodologies, which offer different approaches for dealing with linked data in specific areas and for specific purposes.
In the section 4.2 we have explained only those which passed the three filters explained in the section 4.1, but there are more, in several application areas and with different architectures and other specifications.
[8],[18],[6],[5] proposed eventually approaches to produce linked data but are generally specific to a use case and can hardly be adapted to other use-cases.
In[22],[30] and [10], the authors presented methodologies for linking data in the media and library domains respectively. [20] in finance, [24] in education for the Department of Computer Science at the University of Sheffield,UK, [26] in healthcare...and many others.

As for the conversion step, we can notice that most of them share the same steps but use different techniques to achieve the task. The difference between approaches is the way the mapping is done: for example [16] and [10] create RDF resources then map them to resources from existing vocabularies while [17] and [14] maps directly to resources of existing vocabularies. A special case is [1] who used preset RDF templates and fill in the data from files since all incoming files have same format.
Some approaches proposed additional steps to enhance the quality of data produced: [17] proposed to detect links between columns to create new RDF classes for a better precision while [25] worked on algorithms to detect the data types of cells (literal or object URIs).
[10] proposed an approach that is totally different: it converts data to a relational database before transforming it to RDF format.

5 Conclusion

In this paper, we have analyzed several approaches to transform open data to linked data, then compared them and classified them according to different criteria and identified the most significant limitations they face.

The study of the existing linked data approaches provided us with a practical analysis of the methods and techniques that can be used in each step of the life-cycle for a linked data dataset. We drew experience on the different ways to get the data, different modeling approaches, the best ontology (re)use practices, the various methods for data transformation which, the different ways of publishing the dataset on the web, as well as the extensive ways the datasets can be used in real-world applications and services. We identified the specifics, advantages, and drawbacks of the existing linked data methodologies.
Our future efforts will be guided by the perspective of the semantic web realization through linked data principles. We plan to propose a generic and fully automatic transformation approach that supports more than one entry file type and provides links to external datasets.

References

1. Ghislain Atemezing, Oscar Corcho, Daniel Garijo, José Mora, María Poveda-Villalón, Pablo Rozas, Daniel Vila-Suero, and Boris Villazón-Terrazas. Transforming meteorological data into linked data. *Semantic Web*, 4(3):285–290, 2013.
2. Sören Auer, Lorenz Bühmann, Christian Dirschl, Orri Erling, Michael Hausenblas, Robert Isele, Jens Lehmann, Michael Martin, Pablo N Mendes, Bert Van Nuffelen, et al. Managing the life-cycle of linked data with the lod2 stack. In *International semantic Web conference*, pages 1–16. Springer, 2012.
3. Stefan Bischof, Axel Polleres, and Simon Sperl. City data pipeline. *Proc. of the I-SEMANTICS*, pages 45–49, 2013.
4. Christian Bizer, Tom Heath, and Tim Berners-Lee. Linked data: The story so far. In *Semantic services, interoperability and web applications: emerging concepts*, pages 205–227. IGI Global, 2011.
5. Stéphane Corlosquet, Renaud Delbru, Tim Clark, Axel Polleres, and Stefan Decker. Produce and consume linked data with drupal! In *International Semantic Web Conference*, pages 763–778. Springer, 2009.
6. Ben De Meester, Wouter Maroy, Anastasia Dimou, Ruben Verborgh, and Erik Mannens. Declarative data transformations for linked data generation: the case of dbpedia. In *European Semantic Web Conference*, pages 33–48. Springer, 2017.
7. Tore Dyba, Torgeir Dingsoyr, and Geir K Hanssen. Applying systematic reviews to diverse study types: An experience report. In *First International Symposium on Empirical Software Engineering and Measurement (ESEM 2007)*, pages 225–234. IEEE, 2007.
8. Mamdouh Farouk and Mitsuru Ishizuka. Converting db to rdf with additional defined rules. In *WEBIST*, pages 709–716, 2012.
9. Archit Gupta, Krishnamurthy Koduvayur Viswanathan, Anupam Joshi, Timothy Finin, and Ponnurangam Kumaraguru. Integrating linked open data with unstructured text for intelligence gathering tasks. In *Proceedings of the 8th International Workshop on Information Integration on the Web: in conjunction with WWW 2011*, pages 1–6, 2011.
10. María Hallo, Sergio Luján-Mora, Juan Carlos Trujillo Mondéjar, et al. Transforming library catalogs into linked data. 2014.
11. Tom Heath and Christian Bizer. Linked data: Evolving the web into a global data space. *Synthesis lectures on the semantic web: theory and technology*, 1(1):1–136, 2011.

12. Pascal Hitzler, Markus Krotzsch, and Sebastian Rudolph. *Foundations of semantic web technologies*. Chapman and Hall/CRC, 2009.
13. Bernadette Hyland and David Wood. The joy of data-a cookbook for publishing linked government data on the web. In *Linking government data*, pages 3–26. Springer, 2011.
14. Gaurav Jaglan and Saniay Kumar Malik. Lod: Linking and querying shared data on web. In *2018 8th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, pages 568–573. IEEE, 2018.
15. Barbara Kitchenham. Procedures for performing systematic reviews. *Keele, UK, Keele University*, 33(2004):1–26, 2004.
16. Timothy Lebo and Gregory Todd Williams. Converting governmental datasets into linked data. In *Proceedings of the 6th International Conference on Semantic Systems*, pages 1–3, 2010.
17. Varish Mulwad, Tim Finin, and Anupam Joshi. Automatically generating government linked data from tables. *UMBC Faculty Collection*, 2011.
18. Axel-Cyrille Ngonga Ngomo, Sören Auer, Jens Lehmann, and Amrapali Zaveri. Introduction to linked data and its lifecycle on the web. In *Reasoning Web International Summer School*, pages 1–99. Springer, 2014.
19. Noorafiza Norman and Nurul Hawani Idris. Exploring a linked data approach in accessing physical oceanography archive. In *2017 IEEE 15th Student Conference on Research and Development (SCOReD)*, pages 61–66. IEEE, 2017.
20. Seán O'Riain, Edward Curry, and Andreas Harth. Xbrl and open data for global financial ecosystems: A linked data approach. *International Journal of Accounting Information Systems*, 13(2):141–162, 2012.
21. Zuzana Pelikánová. Google knowledge graph. 2014.
22. José Luis Redondo-García, Vicente Botón-Fernández, and Adolfo Lozano-Tello. Linked data methodologies for managing information about television content: Applying linked data principles in the ontotv system, in order to improve the collection processes and the way television information is accessed. In *7th Iberian Conference on Information Systems and Technologies (CISTI 2012)*, pages 1–6. IEEE, 2012.
23. Octavian Rinciog and Vlad Posea. Govlod: Towards a linked open data portal. In *International Semantic Web Conference (Posters & Demos)*, 2016.
24. Matthew Rowe and Fabio Ciravegna. Data. dcs: Converting legacy data into linked data. *LDOW*, 628, 2010.
25. Kumar Sharma, Ujjal Marjit, and Utpal Biswas. Automatically converting tabular data to rdf: An ontological approach. *Int J Web Semant Technol*, 2015.
26. Cui Tao, Dezhao Song, Deepak Sharma, and Christopher G Chute. Semantator: Semantic annotator for converting biomedical text to linked data. *Journal of biomedical informatics*, 46(5):882–893, 2013.
27. Boris Villazón-Terrazas, Luis M Vilches-Blázquez, Oscar Corcho, and Asunción Gómez-Pérez. Methodological guidelines for publishing government linked data. In *Linking government data*, pages 27–49. Springer, 2011.
28. David Wood, Marsha Zaidman, Luke Ruth, and Michael Hausenblas. *Linked Data*. Manning Publications Co., 2014.
29. Leila Zemmouchi-Ghomari. *Linked Data, a manner to realize the web of data*, pages 87,113. 04 2018.
30. Dydimus Zengenene, Vittore Casarosa, and Carlo Meghini. Towards a methodology for publishing library linked data. In *Italian Research Conference on Digital Libraries*, pages 81–92. Springer, 2013.