Semantic Web 0 (0) 1
IOS Press

Methodologies for publishing linked open government data for the data on the web: a systematic mapping

Bruno Elias Penteado *, Seiji Isotani
Institute of Mathematics and Computer Science, University of São Paulo, SP, Brazil
E-mail: brunopenteado@usp.br

Abstract. Since the beginning of the release of open data by many countries, different methodologies for publishing linked data have been proposed. However, they seem to not be adopted by early studies exploring linked data, for different reasons. In this work, we conducted a systematic mapping in the literature with the aim of synthesizing the different approaches around the following topics: common steps, associated tools and practices, quality assessment validations and evaluation of the methodology. The findings show a core set of activities, based on the linked data principles, but with very important additional steps for practical use in scale. Although a fair amount of quality issues are reported in the literature, very few of these methodologies embed validation steps in their process. We describe an integrated overview of the different activities and how they can be executed with appropriate tools. We also present research challenges that need to be addressed in future works in this area.

Keywords: linked data, linked open data, linked open government data, systematic mapping, methodologies

1. Introduction

As OGD information have proliferated in this last decade, linking and combining datasets have become one of the major topics for the data consumers. The pioneering initiatives in the U.S. and U.K. to produce linked government data have shown that creating high quality linked data from raw data files requires considerable investment into reverse-engineering, documenting data elements, data clean-up, schema mapping, and instance matching [1, 2]. A bulk of data files were converted using simple algorithms, but without much curation efforts, limiting the practical value of the resulting RDF. Alternatively, datasets which are curated and of high quality are limited to restricted subjects, due to the needed effort to create these datasets. The very few public data initiatives that do follow the Linked Data (W3C on Linked Data 2016) paradigm mostly focus only on the metadata for the discovery layer of the datasets, therefore leaving the significant value of analyzing and linking the actual information contained in the data itself by large unexploited, lacking practical approaches for publishing high quality linked government data [1, 3]. According to a report from the World Wide Web Foundation [4], only 7% of the data is fully open, only half of the datasets are machine-readable and only one fourth has an open license. This same reports argues that OGD data needs sustained political will both from a management perspective (process, responsibilities, timelines, etc.) and from data management perspective (guidelines for metadata and publication frequency, documentation, quality assurance processes, user feedback, among other points).

The production of linked data has been increasing since its conception, as can be seen from the LOD Cloud [5], and compiled in Figure 1. Government data has many important applications [6] and it is one of the most popular domain categories of the LOD cloud, published in Datahub[1], with almost 200 linked datasets to date.

*Corresponding author. E-mail: brunopenteado@usp.br.

[1] http://datahub.io

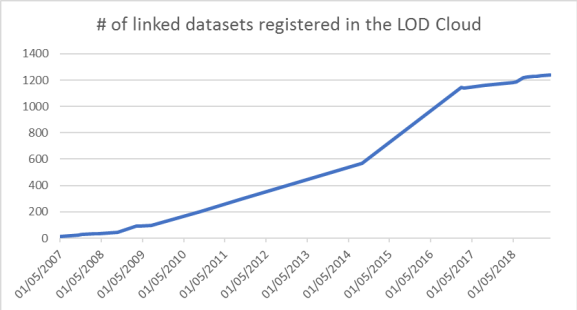

Fig. 1. Number of datasets in the LOD cloud, since 2007 (numbers taken from http://lod-cloud.net.)

Even though Semantic Web technologies based on this idea have flourished, until recently only a small portion of the information on the World Wide Web is presented in a machine-comprehensible way. Particularly, in open government data, this number is still very low in comparison to other levels of open data (CSV, XLS and XML files, in most cases). In [7], the authors elicited open datasets from federal, state and municipality-level in Brazil, and encountered no files with linked data and just one case in which RDF datasets were found. A similar picture in Colombia [8], Italy [9] and in Greece [10], with 5%, 5% and 2% of the datasets in the 4th or 5th level, respectively. A look into the *data.gov* portal (from the US, with different national levels), shows that there is around 2.5% of datasets in RDF format [2], not explicit if they are in the 4th or 5th level. This may be due to the fact that government initiatives are evaluated according to whether they comply or not with the law, and not based on the usefulness of the information provided [11].

As will be outlined in the next section, some methodologies for publication of linked open government data were proposed, but the adopters claim that they are too generic for their purpose, without guidelines for software tools, templates, techniques or other artifacts that could help in the adoption of this technology [9, 12, 13]. As an effect, although there exist many guidelines for publishing linked data on the Web, many producers do not have sufficient knowledge of these practices, having few studies detailing the whole process, leaving out the methods, tools, and procedures used [14], and proposing ad-hoc methods to produce linked open data, usually based only on the 4 principles with different interpretations on how to implement them. In [15] it is indicated, based on interaction with practitioners, that literature on publishing LOGD has dealt with less complex, non-operational datasets and needs an engineering point of view, the identification of practical challenges and consider the organizational limitations. In [9] the authors also argue on similar issues, such as linking quality to external datasets, the lack of domain-specific ontologies and their proper alignment when they exist and the expertise in SPARQL queries when consuming linked data.

In addition, several problems have been occurring regarding the quality of the linked data published on the Web. For instance, [16] identified 3 recurrent problems by surveying LOD papers from the Semantic Web Journal: the existence of inadequate links in the published dataset, compromised quality of the dataset and global impact of the LOD dataset in terms of replicability of the overall process.

In this work, we aim to make a systematic mapping of the literature regarding the processes and methodologies developed to publish linked open government data on the Web and discuss what key challenges remain to be explored.

2. Background

2.1. Open government

Since the late 2000's governments around the world started to move towards publishing increasing volumes of government data on the web, perhaps most notably after the launch of national data portals in the United States (www.data.gov) and the United Kingdom (www.data.gov.uk). This opening has been happening according to the Open Data philosophy[3], making government data freely available to everyone without any restriction. Since then, many countries and cities started to publish their information on the web. The main motivations for such movement was the expected impact in society: increasing transparency and democratic accountability, supporting economic growth by stimulating new data-based products and services, and improving how public services are delivered [1, 17]. As a result, citizens that search open data

[2] Although RDF is not the only serialization format towards linked data, it is acknowledged that it is the most popular format and can be used here as a proxy for the use of linked open government data.

[3] Open data refers to data that "can be freely used, reused and redistributed by anyone". Definition available at: http://opendatahandbook.org/guide/en/what-is-open-data/

government (OGD) on the Web are involved in a time-consuming process, which includes: (a) identification of relevant sources, (b) consistency checking of information (c) aggregation of information.

OGD provision presents some limitations that hamper data reuse. The organizational limitations originate mainly from the fact that in public administration each agency manages data according to its mandate, since there is no central entity assigned with this role. In addition, public agencies formulate hierarchical structures that contain a number of administrative levels. This organizational structure of the public sector suggests that in certain cases public agencies in different administration levels and different functional areas produce, maintain and possibly disseminate similar data, i.e. data about the same real-world object (e.g. a specific school) or the same real-world class (e.g. schools) [18].

Many studies [19–21] illustrate that the use of OGD is often hampered by the multitude of different data formats and the lack of machine-readable data, imposing restrictions on their consumption by end-users, in terms of discoverability, usability, understandability, access, and quality, among other aspects. Although publishing government information as open data is a necessary step to realize the mentioned benefits, it is not sufficient. In practice, gaining access to raw data, placing it into a meaningful context, and extracting valuable information is extremely difficult [22]. A possibility of reusing open government data is by linking them to other data, so that relationships with other data can be explored [23].

2.2. *Linked data principles*

In summary, Linked Data is about using the Web to create typed links between data from different sources - with diverse combinations of organizations, data formats and exchange standards [24]. It refers to data published on the Web in such a way that it is machine-readable, its meaning is explicitly defined, it is linked from/to other external data sets. Berners-Lee [23] outlined a set of design principles for publishing and connecting data on the Web, to become part of a single global data space, establishing the principles for linked data:

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.

These were the initial principles to publish linked data on the Web. Berners-Lee [23] extends these principles to include the concept of open, by defining the 5-star scheme for linked open data, interested particularly in government data, but arguing that it could be also used for other types of sources:

★ Available on the web (whatever format) but with an open license, to be Open Data
★ ★ Available as machine-readable structured data (e.g. excel instead of image scan of a table)
★ ★ ★ as (2) plus non-proprietary format (e.g. CSV instead of excel)
★ ★ ★ ★ All the above plus, use open standards from W3C (RDF and SPARQL) to identify things, so that people can point at your stuff
★ ★ ★ ★ ★ All the above, plus: link your data to other people's data to provide context

Linked data extends the concept of *open data*. Open data is data that is publicly accessible via the Internet, without any physical or virtual barriers to accessing them. Linked Data, in turn, is data that allows relationships to be expressed among these data, enriching the datasets with complementary information from elsewhere [25]. This extension carries complex issues such as different granularities, data alignment, transformation, and storage but also brings important benefits: contextualization of data and background information, by using additional information from other sources; automatic reasoning by software agents, among others. The emergence of the Linked Data principles has introduced new ways to integrate and consolidate data from various and distributed sources [24, 26]. This 'Web of data' aims at replacing isolated data 'silos' with a giant distributed dataset built on top of the Web architecture, usable both by software agents and humans [27, 28].

3. Related works

The production and publication of linked data are intensive engineering processes that demand high efforts to achieve high quality and existing general guidelines may not be sufficient to make the processes repeatable [29]. Since the conception of linked data, some principles and processes were proposed, with varying degrees of sophistication, practices, and tools.

The following studies presented some form of synthesis from previous methodologies. In [13] linked data publishing methodologies are elicited, mostly from the government domain, in particular, three studies which were adopted by the W3C Government Linked Data Working Group to create guidelines for accessing open government data using the Linked Data principles [30]. In their conclusion, the authors argue that these methodologies, although valuable as guidelines, either do not consider important steps of the linked data production lifecycle or are described too generically so that publishers have to figure out the tools needed to accomplish properly the publishing process as a whole. In [11] the authors also presented a systematic review of OGD initiatives (not linked data) and presented a lifecycle deduced from the related papers, along with related challenges in different levels (organizational, economic and financial, policy, legal and cultural). In [31] the authors compiled the steps from 8 different linked open data methodologies but did not specify what were the criteria to select the primary studies. However, the proposed framework is also in a high level of abstraction. The LOD2 Project [32] also developed a lifecycle for linked data and provided software tools for the steps, although leaving out important steps - such as data modeling, alignment and the publication of the data on the Web.

This study complements other systematic mappings or reviews, such as those of [6], which surveyed the adoption of best practices for publishing linked data, discussing which of the W3C best practices [30] are explicitly more present in the literature; and the systematic review on the use of software tools for linked data publishing, conducted by [33], which points out that most of the current state-of-the-art tools are concentrated in only a few of the steps of the publishing process, leaving important steps out. These systematic mappings did not provide information on the tasks involved during the process of linked data production. In addition, in [14] the authors performed a systematic mapping of publishing and consuming data on the web; thus a more generic approach than in the present study. One of their findings was that most of the papers surveyed did not mention publishing methodologies (28 out from 46) and most of the ones which did (12 from the remaining 18) just used the basic linked data principles as a guideline for the process. Other systematic mappings/reviews were carried out in different domains, such as enterprise linked data [34] and education [35, 36], and applications such as linked data mashups [37], recommender systems [38], quality assessment [39].

To the best of our knowledge, there is no systematic mapping of linked open government data methodologies in the literature. Thus, in this work, we sought to make a systematic mapping of methodologies proposed in the literature, in order to provide a synthetic comparison of the steps, tools, and validations proposed by these methodologies and how they were evaluated, so to map how such methodologies can be extended for further reuse.

4. Methodology

In this paper, we use the systematic mapping method [40], aimed to identify research related to a specific topic to answer a broad question, essentially exploratory (e.g. *What is known about X?*), preserving the reproducibility of the study - since the objective of this paper is to present an overview of the literature to investigate the development of methodologies for publishing linked open government data. This is a complementary perspective with the systematic review [41] in which the effectiveness of treatments are aggregated and compared. The systematic mapping consists of 5 steps: definition of the research questions, search for primary studies, screening of papers for inclusion and exclusion, keywording of abstracts and data extraction and mapping of studies. The complete results are available online: https://bit.ly/319BGAH.

4.1. Research questions

The research questions defined in this work aim to gather information about how to effectively publish linked open data in government settings, both for the steps involved and for the tools developed to accomplish it. We argue that this is an important contribution to the scientific community and practitioners alike, to describe what has been done and the gaps that should be addressed to a better outcome of LOGD policies. To that end, we defined the following research questions:

- *RQ1. What are the common steps among the different methodologies proposed?*
- *RQ2. How were the methodologies evaluated empirically?*

– *RQ3. What tools and vocabularies were used or recommended to support the steps?*

– *RQ4. What kinds of validations were employed to assure better data quality practices?*

The answers to these questions provide a big picture of the relevant literature, with important steps to suggest a clear methodological framework for the publication of LOGD.

4.2. Search strategy

The following datasets were used for this systematic mapping, since they are the most significant repositories in subjects that involve Computer Science: ACM Digital Library, IEEE Explore, Science Direct, Springer Link, ISI Web of Knowledge and Scopus.

The identified keywords are *methodology*, *publishing* and *linked open government data*, which were grouped these terms and their synonyms were considered to elaborate on the search string (Table 1).

Table 1
Terms used for the search.

(*methodology* OR *process* OR *pipeline* OR *guideline* OR *"best practices"* OR *framework*)
(*publishing* OR *publication* OR *production* OR *opening*)
(*"linked government data"* OR *"linked open government data"* OR *"government linked data"* OR *"government open linked data"* OR (*"open government"* AND (*"linked data"* OR *"linked open data"*)))

4.3. Study selection

The selection of the studies should reflect the primary works with the purpose of identifying different types of methods used to publish linked open government data. To that end, we elaborated the following criteria:

– Inclusion criteria
  * The study provides a process for publishing linked data in government settings as the main contribution;
  * The study is from a peer-reviewed vehicle;
  * The language of the study is English;
  * The text of the study is available;

– Exclusion criteria
  * The study does not present a process for publishing LOGD;
  * The study is a previous version from another in the list;
  * The study focus on the application of LD in a specific domain;
  * The study only investigates one step of the process;
  * The study does not investigate linked data, but open data more generally.

The procedure for selecting the primary studies for this mapping was carried out in late April/early May 2019. In cases where multiple papers from the same authors referred to the same topic, but in different stages of maturity, we chose the most recent one - obviously, if related to the topic under study. As shown in Figure 2, we applied the concrete string, adapted according to each database, separately. Subsequently, the duplicated versions were removed. Next, we applied the inclusion and exclusion criteria to every study, considering the title, abstract and, when in doubt, the full-text. Finally, we had the final set which we could extract the data to answer the research questions.

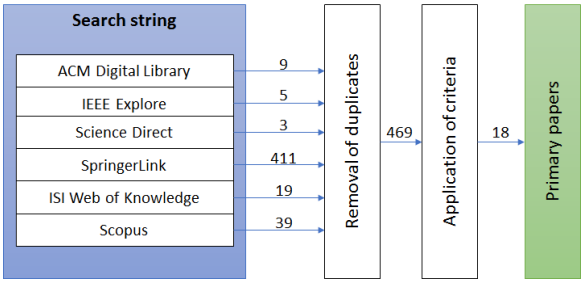

Fig. 2. Procedure to select the final studies.

4.4. Threats to validity

Systematic mappings may present multiple threats to validity [42]. We composed the search string into three aspects: process, publishing and linked open government data. The use of synonyms was based on textual analysis. These terms, particularly for linked open government data, were difficult to specify, because they had different ordering of words and sometimes not used together. In order to control the quality of the results, we used the studies described in the W3C Linked Data Best Practices [30] as a control to tune the query string. We also restricted ourselves to the exe-

cution of the query in the data repositories, not applying manual searches in other platforms. Some papers were not available, and for those, we searched on the web for a copy and contacted the first author to try to obtain a copy of the work, but sometimes that was not possible.

5. Results

The final selection resulted in 18 primary papers, with dates ranging from 2011 to 2019, which were used to extract information regarding the research questions. Table 2 presents the selected papers.

Table 2
Final set of primary papers selected.

ID	Reference	Publication	Year
W1	Laessig et al. (2019) [12]	Chapter	2019
W2	Martins et al. (2018) [43]	Conference	2018
W3	Fleiner (2018) [44]	Conference	2018
W4	Krataithong et al. (2018) [45]	Conference	2018
W5	Elmekki et al. (2018) [46]	Conference	2018
W6	Buranarach et al. (2017) [47]	Conference	2017
W7	Klein et al. (2016) [48]	Conference	2016
W8	Ngomo et al. (2014) [49]	Chapter	2014
W9	Sorrentino et al. (2013) [50]	Chapter	2013
W10	Kaschesky & Selmi (2013) [51]	Conference	2013
W11	Al-Khalifa (2013) [52]	Conference	2013
W12	Janev et al. (2012) [53]	Conference	2012
W13	Maali et al. (2012) [2]	Conference	2012
W14	Hyland & Hyland-Wood (2011) [54]	Chapter	2011
W15	Villazón-Terrazas et al. (2011) [55]	Chapter	2011
W16	Cifuentes-Silva et al. (2011) [56]	Conference	2011
W17	Salas et al. (2011) [57]	Chapter	2011
W18	Lebo et al. (2011) [58]	Journal	2011

RQ1. What are the common steps among the different methodologies proposed?

This research question aimed to map what are the commonalities and differences among the different methodologies that have been proposed for publishing linked open government data. One first challenge was to find the correct granularity for this. Most of the studies divided the tasks of publishing into phases and, in turn, in more atomic steps with clearer outputs. To analyze these data, we mapped out all the activities that were explicitly described as an important step in the papers, creating a matrix of steps *x* studies, as in Figure 3.

Figure 3 lists all the explicit tasks identified and close to their ordering, as described in the papers. The first step, sometimes implicit, concerns the selection of datasets to be linked and consider to leverage existing open datasets or to expose new ones, the identification of their structure, and so on. Next, some studies consider cleaning up the data, to remove inconsistencies, typos, or problems with the structure of the data. As one of the pillars of linked data is dereferenciability using HTTP URIs, the careful design of URIs were also considered. Another step is the definition of vocabularies[4], again analyzing when to reuse existing ones or to build new ones, depending on the context of the data. The specification of metadata - both for the dataset and the data content - is also considered as a step to describe what is being published to the potential consumers. Next, the careful mapping of the vocabularies to data is performed. In addition, there must be indicated the external sources with which one wants to link the data, such as Dbpedia or GeoNames, and the step of creating these links must be carried out. Some studies use this linking to perform the enrichment of the dataset, importing data from the external sources. With all this material, the step of transforming 'raw' data into RDF is executed. The conversion of formats depends on manual mapping or inferred structures from data and may need some cleanup tasks too. Since datasets and their distributions change over time, a mechanism for keeping track of versions are also needed. With these data linked, one needs to publish them online, to be reused in data portals or SPARQL endpoints. That step can be leveraged by publicizing it in different open data indexes and facilitating for search engines or engaging with the community of users and consumers. Some studies point to the importance of creating applications with the data, to help the community raise its awareness of it. With all set, it is important to have a plan to keep all this working over time. To that end, the studies specify tasks to maintain the data portal and to define non-functional requirements (performance levels on serving data, uptime, security profiles to access the data, and so on) and maintenance tasks (e.g. checks of data availability).

[4] In most of the studies the following terms are used interchangeably: vocabularies, taxonomies and ontologies. In this work we use the same approach.

Step / Article	W1	W2	W3	W4	W5	W6	W7	W8	W9	W10	W11	W12	W13	W14	W15	W16	W17	W18	#
Select data	X	X	X	X	X	X	X	X		X	X	X	X		X	X		X	15
Clean-up source data	X				X	X	X						X					X	6
Design URIs			X											X	X	X		X	5
Define vocabularies	X	X	X	X						X	X	X		X		X			9
Specify metadata	X	X	X									X		X	X	X		X	8
Map vocabularies to data		X		X		X	X	X	X	X					X		X		9
Link to other data sources	X		X		X			X	X	X		X	X		X				9
Enrich the datasets							X	X	X	X		X						X	6
Convert to RDF	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	18
Clean-up RDF data															X				1
Version the datasets																		X	1
Define licenses	X														X	X			3
Publish the data	X	X	X	X		X	X	X	X	X	X	X	X	X	X	X		X	16
Enable discovery			X							X		X			X				4
Build/reuse apps on top of data			X				X	X		X					X	X			6
Create and maintain the data portal												X				X			2
Engage with the community	X																		1
Define non functional requirements								X		X						X			3
Define maintenance tasks								X								X			2

Fig. 3. Mapping of steps and studies for the selected methodologies. The last column accounts for the total number of appearances.

Observing the last column, we can see that some steps are much more present than others. In particular, the very basic steps are: *select the data*, *convert to RDF* and *publish the data*. Next, *defining and mapping vocabularies*. In fact, some studies are closely limited to these core steps [e.g. *W6*, *W11*, *W13*], reaching the 4th level of linked data. Next, the *interlinking of different data sources*, configuring the 5th level. A similar number of appearances for the metadata specification. There are more sparse activities, considering the publishing process in a broader sense, such as versioning and licensing of data, community feedback, and maintenance tasks.

RQ2. How were the methodologies evaluated empirically? In this work, we consider the methodologies for publishing linked open government data as artifacts designed to solve problems of a particular domain, achieving knowledge and understanding of it, as conceptualized in the Design Science Research field [59]. Thus, we used the categorization of [60], derived from works in design science research, to classify the different evaluation methods applied in the selected studies.

As illustrated in Table 3, 39% of the studies (7 out of 18) did not provide an empirical evaluation in the paper, being restricted to make a list of steps and recommendations, mostly justified by the basic principles of linked data and the 5-stars schema. Most of the papers (9 out of 18) provided illustrative scenarios of the application of the methodology. The actual validations were varied, ranging from the visualization of weather statistics [*W4*] to cataloging a national library [*W16*], and also batches of government data [e.g. *W6*, *W18*]. According to this evaluation framework, illus-

Table 3
Evaluation methods adopted in the selected studies.

Evaluation method	Description	Studies
Illustrative scenario	Application of the method in real-world data aimed at illustrating the utility of the artifact	[W4, W6, W7, W9, W12, W13, W15, W16, W18]
Prototype	Implementation of an artifact aimed at demonstrating the utility or suitability of the artifact	[W2, W17]
Logical argument	An argument with face validity; without empirical experimentation	[W1, W3, W5, W8, W10, W11, W14]

trative scenarios differ from case studies because the latter involves analyzing the impact of the intervention in the natural environment. In the selected studies, there was no article which provided this sort of evaluation, being restricted only to prove the concept. In addition, two of the papers [*W2* and *W17*] focused on showing how a tool could support the process and detailed its features. We must emphasize that in this work we focus on the actual validation explicit in the paper. Some works build on the authors' previous experiences in open government data projects or perhaps on validations carried out in other stages of the research and that could not be retrieved.

RQ3. What tools and vocabularies were used/ recommended to support the steps? Although the prescription of tools is not mandatory in a methodology, it surely offers a good starting point for practitioners, in making decisions like *buy vs. build* approaches. As argued in [25] and [26], working on small datasets is a task that can be tackled manually, for small and static datasets. However, LOD projects, particularly in governments' scope with large and diverse datasets, the use of tools is necessary to ease the effort in an automatic or at least semi-automatic approach. Thus, we consider that this is an important source of information. To that end, we used the same steps identified in *RQ1* and mapped how each of the studies dealt with it in their respective papers. Figure 4 shows the mapping of tools used in the selected studies.

The biggest diversity in tools was found for the data conversion step, with many different tools proposed, used or suggested. In this case, the choice of a particular tool depends on the nature of the data source. For instance, when the data is extracted directly from a relational database, the D2RQ platform was the most used. In other cases, where other data formats are used, such as the common OGD formats of CSV, XML and JSON, OpenRefine was most frequently used. Most of these tools also provide a feature that supports the mapping of vocabularies during the task of converting raw data into RDF - thus the overlapping of tools in both steps.

For the selection of data, some custom tools were built, aiming to extract data from a repository and use it as input for the next steps - assuming that the datasets were already specified and are available (either in a relational database or an open data catalog). For the definition of vocabularies two distinct approaches were identified: tools to search for existing vocabularies (such as LOV, Swoogle) and tools to create new vocabularies, like Protégé, OntoWiki, and TopBraid Composer. For cleaning up the data, OpenRefine was used in two studies, along with two other custom tools. For the design of URIs no tools were used, but guidelines, especially the Cool URI guideline[6], which recommends practices on how to model instances using HTTP URIs. Other guidelines were also listed: Designing URI for the UK Public Sector[7] and Style Guidelines for Naming and Labeling Ontologies[8].

Concerning the storage of linked data, two tools were most used: CKAN and OpenLink Virtuoso. CKAN is currently the open data catalog most used by open government initiatives, hosting files and serving them through the Web and their metadata by API interfaces. On the other hand, Virtuoso, an open-source platform used to host RDF knowledge graphs and making them available through a SPARQL interface. Although it provides more flexibility, it also carries problems of usability by end-users - who must be knowledgeable in SPARQL queries - and performance issues, because of the dynamicity of the query results. [9]

RQ4. What kinds of validations were employed to assure better data quality practices?

In software engineering, *verification and validation* are the processes of checking whether a software product meets specifications and that it fulfills its intended purposes. Publishing open government data on the web is a major step, but their value is only as important as

[6] https://www.w3.org/TR/2008/NOTE-cooluris-20081203/
[7] https://www.gov.uk/government/publications/designing-uri-sets-for-the-uk-public-sector
[8] http://dcpapers.dublincore.org/pubs/article/view/3626
[9] Given the number of tools, the list of references can be found in the full report, in appendix: https://bit.ly/319BGAH

Step	Tool/Guideline
Select data	Open Data Kit [W1], KoBo Toolki[W1], UnBGOLD [W2], dcat browser [W13], custom tools [W5,W6]
Clean-up source data	Open Refine [W7, W13], custom tools [W5, W6]
Define vocabularies/ontologies	LOV [W3, W15], Protégé[W3, W15], Semantic Media Wiki [W8], OntoWiki [W8], Swoogle [W15, W16], SchemaWeb [W15], SchemaCache [W15], Neologism [W15], NeOn Toolkit [W15], TopBraid Composer [W15], Altova [W15]
Design URIs	Cool URIs [W3, W14, W15, W16], Designing URIs for the UK Public Sector, Style Guidelines for Ontologies
Specify metadata	UnBGOLD [W2], VoID [W3, W15, W18], Open Provenance Model [W15, W18], Dublin Core [W3], Provenir [W18]
Map vocabularies to data	OpenRefine [W7, W10, W13, W15], D2RQ [W4, W6, W9, W15], UnBGOLD [W2], Semantic Media Wiki [W8], OntoWiki [W8], WebDAV [W10], Sponger [W10], StdTrip [W17], XLWrap [W15], RDF123 [W15], NOR20 [W15], UltraWrap [W15], GRDDL [W15], TopBraid Composer [W15], ReDeFer [W15], any23 [W15], Stats2RDF [W15]
Link to other data sources	SILK [W3, W8, W9, W10, W15], LIMES [W3, W8, W15], OpenRefine [W10, W13], RKBExplorer [W8], GNAT [W8], RDF-AI [W8], Pundit [W10]
Enrich the dataset	OpenRefine [W10], MOMIS [W9], Fusepool P3 [W7], csv2rdf4lod [W18], DL-Learner [W8], Protégé reasoners [W8], OntoWiki plugins [W8]
Convert to RDF	OpenRefine [W7, W10, W13, W15], D2RQ [W6, W8, W9, W15], Jena [W2, W4], Kettle [W16], StdTrip [W17], csv2rdf4lod[W18], OpenCalais [W8], Alchemy [W8], FOX [W8], Sparqlify [W8], Virtuoso RDF Views [W8], WebDAV [W10], Sponger [W10], XLWrap [W15], RDF123 [W15], NOR20 [W15], UltraWrap [W15], GRDDL [W15], TopBraid Composer [W15], ReDeFer [W15], any23 [W15], Stats2RDF [W15], custom tools [W5, W11, W12]
Clean-up RDF data	RDF Alerts [W15], sameAs Link Validator [W15]
Version the datasets	-
Define licenses	Creative Commons Choose [W1]
Publish the data	CKAN [W1, W2, W12, W13, W14], Virtuoso [W6, W10, W15, W16], Socrata [W1], OpenDataSoft [W1], Jackan [W2], OAM Framework [W4], LMF [W9], Clerezza [W10], Jena TDB [W10, W15],Fuseki [W13], Sesame [W15, W16], 4Store [W15, W16], OWLIM [W15, W16], YARS [W15], Redland [W16], Bigdata [W16]
Enable discovery	Sitemap protocol [W3, W15, W16], Datahub [W3], CKAN.net [W13, W15]
Build apps on top of data	Pubby [W15, W16], LodLive.it [W10], Relfinder [W10], D3JS [W10], Elda [W16], D2R Server [W16], djubby [W16], WESO DESH [W16]
Create and maintain the data portal	CKAN [W12]
Engage with the community	-
Define non functional requirements	-
Define maintenance tasks	Trelis [W8], ProLOD [W8], LinkQA [W8], WIQA [W8], Sieve [W8], tSPARQL [W8], CTIC Vapour [W16], RDF/XML Validator [W16]

Fig. 4. Artifacts used or suggested by the studies, according to the steps previously identified[5]

their quality. As a complex process, verification and validation tasks could be used to better guarantee the quality of the data produced. Despite data quality in LOD being an essential concept, the autonomy and openness of the information providers make the Web vulnerable to missing, inaccurate, incomplete, inconsistent or outdated information [Ngomo et al., 2014]. And, as argued previously, even with all the effort made the final result may not reach a high quality. Thus, we sought to search which validation tasks were employed by the studies during the process.

Few studies proposed an explicit phase or steps to make validations throughout the lifecycle of linked data production. *W8* brings the most detailed tasks, with a phase dedicated to linked data quality and its respective validations. The authors considered the work of Zaveri et al. [39] and listed 18 quality dimensions and 68 metrics, divided into 4 groups: accessibility, intrinsic, contextual and representational. However, the study did not apply it in a real case study, only with an illustrative example. *W15* employed two validations in their methodology: in the data clean-up phase, to check for RDF, accessibility, vocabulary, and data types mistakes or errors; and in the final of the linking phase, in which domain experts should revise the automatic links created with tools like SILK or LIMES. *W10* provides a validation phase, between the linking phase and the release of the data. In this phase, the authors claim that data should be checked for accuracy, accessibility, consistency, completeness, visibility, cataloging, promotion, compliance and privacy. The study does not detail this phase nor apply it in a case study.

W6 presents a step for validating tabular data, after the automatic collection from a digital catalog and before converting them to RDF. The authors presented their algorithm and applied it in datasets from Thailand open data portal, evaluating the precision and recall metrics of the algorithm for identifying structural problems.

Other studies mention the importance of validations during the process. However, they offered suggestions and did not contemplate dedicated tasks. *W16* performed in their case study the validation of the RDF conversion for the correctness of format. *W7* also did quality checks after the data transformation, and only did it in the case studies and without further detailing. *W5* discusses the problem of data incompleteness but does not detail how their methodology and architectural components dealt with it. *W4* states that, in their approach, only well-formed datasets could be processed, but did not show how it could be checked in their methodology. *W3* points to the importance of validating the links to external datasets, that should be performed by domain experts. *W1* highlights the importance that potential users need to understand and validate the data, during the data collection phase, however, the authors do not detail how it could be carried out.

6. Discussion

This work sought to make a mapping of methodologies developed for the publication of linked open government data on the Web, given all its specifics. Although the open government data movement is still producing large amounts of data worldwide, the linked data still represents a very small portion of those. Given the distributed nature of the Web and the intrinsic distributed nature of linked data, even if a dataset is correctly deployed, its quality (and therefore, its chance for reuse) depends also on its dependencies on external data, which makes the maintenance of the linked datasets a big challenge.

We notice that important studies were made in the beginning of this decade and it has again been leveraged in the last few years. The reason for the creation of these methodologies in the period of 2011-13 is arguably the deployment of governmental open data portals, such as in the USA (2009) and the UK (2010) that released hundreds of datasets in their first years, glimpsing the opportunity for a Web of data [23]. Many studies in the last few years, returned from the search in the data sources, concerned the application of a method to create linked data for a particular purpose, sometimes based on one of the studies listed here and most of the times by creating an ad-hoc approach for their problems. The justification is mostly that the existing methodologies are too generic and do not consider the particularities of their domain. Some domains were more prevalent in the applications of linked data: geographical data, e-procurement, agricultural and environmental data, smart cities and legislative data. In addition, a subset of the studies investigated just one or a few steps of the whole process, such as techniques for data quality enhancement, automatic interlinking of datasets, vocabularies/ontologies development, the licensing resolution, semantic data extraction from HTML tables, among others.

As pointed in [13], the existing Linked Data methodologies have a varying number of steps, but still generally cover the same activities. The main difference in the methodologies is the grouping of actions within different steps and on different levels of granularity. Apart from some explicit differences, which we will further examine, they cover the palette of actions involved in the process of generating and publishing a linked dataset, and thus can be grouped into six general phases.

Regarding our first research question, we showed the commonalities of the different methodologies. Most of the studies addressed the basic tasks of: selecting data sources, converting them to RDF, linking them to other datasets and publishing the resulting files. Although these are all essential tasks to publish linked data, some of the studies did not mention it explicitly. For example, *W9* used as a starting point a particular dataset from the Italian government, thus not considering the step of selecting data sources and its particular issues. The only task that was explicitly described by all the methodologies was the conversion of OGD data to RDF, rendering all other tasks as auxiliaries to this core activity. However, linked open data is not just transforming tabular data into RDF and putting it on the Web. So, each methodology contributed sparsely with different, yet important, steps that should be considered to achieve a final product with good quality, such as modeling the licenses of the data, the versioning of datasets, the engagement with the community, the definition of non-functional requirements (such as privacy and performance) and important maintenance tasks. Based on the steps extracted from the papers, we built the following process model depicted in Figure 5, with all the steps grouped by the most common phases

present in the studies. It can be used as a roadmap for LOGD initiatives and resource estimation, where managers may decide what level of formalism should be developed according to their context.

As publishing linked data is a very complex process, we argue that these are important aspects that must be taken into account in the scenario of publishing open government data as linked data on the Web. The W3C's recommendation Data on The Web Best Practices [61], although do not focus only on linked data, considers most of these issues and may provide guidance on how to map them into a formal method.

Our second research question assessed how these methodologies were evaluated in their initial proposal. The assessment framework we adopted here was based on the literature of information systems and design science research, which focuses on the design, development and evaluation of artifacts to address real-world problems [59]. The artifact type here is a method, i.e., actionable instructions that are conceptual, not algorithmic. An important phase is the evaluation process, with different degrees of formality. We found that some of the selected papers did not present any formal evaluation of the methodology (logical arguments); mostly written to be used as a tutorial or a set of best practices rather than a formal inquiry. Maybe that is one of the reasons why they are perceived as too generic and not adopted in later works. Two studies (*W2* and *W17*) presented a prototype as the main contribution, embedding their methodology in a software, demonstrating that it works as intended and it is useful for its intended purpose. Thus, we noted a lack of more formal evaluations with the proposed methodologies, in assessing how they modify their context. Although it may not be reasonable to design controlled experiments to evaluate the methodologies, other forms may be employed, such as case studies or action research. According to this framework, both evaluation types investigate how the artifact was used and how it addressed the real-world problem. The illustrative scenarios, on the other hand, applies the artifact to demonstrate its suitability but does not consider how it affected the situation (for instance, the technological impacts or the consumption of the data).

The third research question assessed how these methodologies prescribed tools to support their execution. The use of tools may be considered as a systematic concretization of the methodology since it provides a common ground that can be applied and compared in different situations. As with the second research question, many studies suggested few tools or just a single one to different steps. As they were designed to be generic for different domains, only abstract steps were suggested, leaving open how it can be done in different domains. This may also be a reason why they are perceived as too generic in later works. The major exception in this list was *W8*, which listed lists of tools for every phase that encompasses their methodology, in a 99 pages length report. As with the first question, the bulk of tools were concentrated in the core tasks: the mapping of vocabularies/ontologies to the raw data, the conversion of data files to RDF, and the storage platform (triple stores or open data catalogs). A cross-reference with works such as LOD2 project [32] - developed to provide software stack aiming to support the production of linked data - or OpenGov Intelligence[10], for statistical data, might be useful so that non-expert publishers may become familiar with the whole process and experiment themselves in their context. Other platforms, such as the LinDA project[11] and DataGraft [12] also present a set of tools to deal with the whole process, yet they handle only the most common scenarios.

Our fourth research question explored what validations were employed during the process of linked data production. As pointed previously, data quality is still a big issue for linked open data on the Web, so a validation model throughout the process could bring benefits to the availability of the final product. Few studies presented explicit validation tasks during the process. Most of the studies either just recommended that some steps would be advisable or did not include it at all. The studies which did specify either did not evaluate it with a real case study or did it for specific steps of the process - particularly, to validate the format of the input data (mostly, tabular data) or to validate the links to other datasets identified automatically. The exception was again W8, which provided with a whole phase concerning data quality with many metrics and validations that could be performed in different aspects, but without an actual application. Two studies (*W7* and *W16*) did not prescribe a specific task for validation during the presentation of the methodology but did it in the illustrative scenario that they applied the methodology, what leads to thinking that validations are supposed to be implicit for the entire process. Thus, in accordance with the first research question, the auxiliary steps, along with the core tasks, are important in as-

[10] http://www.opengovintelligence.eu/
[11] http://linda-project.eu/
[12] https://datagraft.io

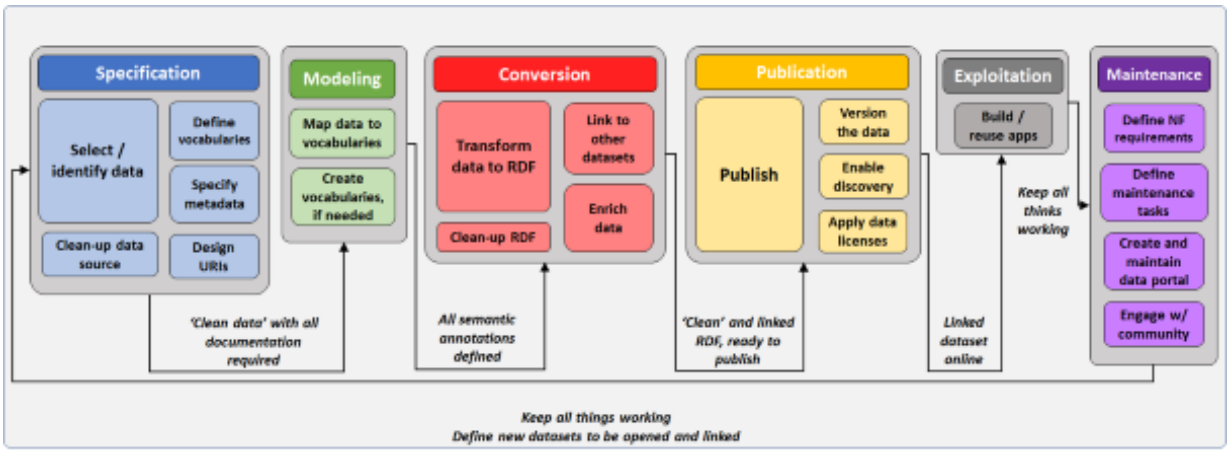

Fig. 5. Process model derived from the steps extracted from the selected papers.

suring a higher quality of the data and should also be considered for validation tasks throughout the process.

7. Research directions

We list in this section some possible research directions concerning improvements in methodologies to publish LOGD, in general. Other important aspects, such as data consumption, are out of the scope of this work.

Considering all the variabilities and commonalities from the different methodologies, we consider helpful to create a process model for publishing LOGD. Since we have core activities, that appears to be common to all the contexts (*RQ1*), it should provide a map so that practitioners could understand the whole picture and make informed decisions on which steps should be used or discarded and their impacts in the final product.

Methodologically, it would be interesting to have longitudinal studies, considering the usage of the linked data, how the methodology evolved in the context in which it was applied, and should also drive for requirements for the maintenance phase. Although illustrative examples are helpful to demonstrate how it can be applied with real data, the production of (linked) open data is a sociotechnical process [19, 62] through which there is a continual interplay between technological (process, tasks, technology) and social aspects (relationships, reward systems, authority structures).

The inclusion of explicit validation steps along the process may be helpful to ensure a higher quality product early on the process. Some validations can be automated, particularly concerning structural aspects and some may be considered to be prone to human analysis, especially in semantic modeling. Methodologies such as the V-model [63] for software development considers a validation point after the end of each phase and could be adapted to this end. Or maybe the application of acceptance criteria for user stories from agile methods. Quality frameworks such as the one provided by Zaveri et al. [39] and the Data on the Web Best Practices [61] could be used to compose these steps.

Another direction is the possibility to make large scale deployment, reusing legacy open data. A large amount of structured and semi-structured data is already available in most of the countries and provide a valuable source to 'cross the chasm' and reach network effects on the already existing data. The task that requires most effort is arguably modeling the data, either by carefully selecting existing and validated vocabularies or by creating new ones, for each of the datasets and their distributions along time. We argue that this could be achieved by deriving ontologies from the data files, from simple automatic mappings [64] to more elaborate approaches [65, 66] as a starting point, leveraging the mature state of the data, applying a pragmatic perspective of linked data [67], which considers ontologies as a lightweight representation tool for an open and decentralized environment like the Web. The evolution of these vocabularies could be done collaboratively by data consumers and domain specialists

inside or outside the government's scope - thus, also decentralized.

As argued previously, the distributed nature of the Web makes it difficult to assure that all linked components are working or have high quality over time. In addition, the lifecycle of governmental datasets is very dynamic, reflecting administrative changes, domain refinement, new legislation or guidelines around the data and so on. Keeping track of these changes and making them transparently available is a big challenge. Thus, the maintenance phase is very important and should be developed further, in order to monitor if what was produced remains valid in this decentralized context.

8. Conclusions

Publishing LOGD is a very complex task. Although the release of OGD is still growing, the steps to transform it to linked data - with high quality - is an open issue. As discussed in this work, there are relatively few linked data on the web and they present quality problems. Although this is a complex multidimensional phenomenon, some technological and methodological approaches may support its development. Some methodologies were carefully designed, but it seems that they failed to base later works on how to publish linked open government data. As argued in [55], there is no one-size-fits-all process and set of tools to publish linked data, given the different contexts, data sources, technologies, etc. However, the products of the process and most of the steps to achieve it are common among different approaches. In this paper, we followed this rationale, by deducing what has been done in different contexts and deriving a unified methodology with practices adopted during the last decade.

References

[1] J. Sheridan and J. Tennison, Linking UK government data, in: *Proceedings of the Linked Data on the Web Workshop (LDOW)*, Raleigh, 2010.

[2] F. Maali, R. Cyganiak and V. Peristers, *A Publishing Pipeline for Linked Government Data*, in: *The Semantic Web: Research and Applications. ESWC 2012*, H. Springer Berlin, ed., 2012. doi:10.1007/978-3-642-30284-8_59.

[3] S. Mouzakitis, D. Papaspyros, M. Petychakis, S. Koussouris, A. Zafeiropoulus, E. Fotopoulou, L. Farid, F. Orlandi, J. Attard and J. Psarras, Challenges and opportunities in renovating public sector information by enabling linked data and analytics, *Information Systems Frontiers* **19** (2017), 321–336. doi:10.1007/s10796-016-9687-1.

[4] W. Foundation, Open Data Barometer, 4th edition, 2017. https://opendatabarometer.org/4thedition/report.

[5] I.C. for Data Analytics, The Linked Open Data Cloud, 2019. http://lod-cloud.net.

[6] D. Feitosa, D. Dermeval, T. Ávila, I.I. Bittencourt, B.F. Lóscio and S. Isotani, A systematic review on the use of best practices for publishing linked data, *Online Information Review* **19**(1) (2018), 107–123. doi:10.1108/OIR-11-2016-0322.

[7] R. Matheus, M. Ribeiro and J. Vaz, Brazil Towards Government 2.0: Strategies for Adopting Open Government Data in National and Subnational Governments, in: *Case Studies in e-Government 2.0*, I. Boughzala, M.S. Janssen and Assar, eds, Springer, Cham, 2014, pp. 1–8. doi:10.1007/978-3-319-08081-9_8.

[8] L.A.R. Rojas, G.M.T. Bermúdez and J.M.C. Lovelle, Open Data and Big Data: A Perspective from Colombia, in: *International Conference on Knowledge Management in Organizations*, L.U. L, O.D. Fuenzaliza, I.H. Ting and D. Liberona, eds, Springer, Cham, 2014, pp. 35–41. doi:10.1007/978-3-319-08618-7_4.

[9] R. Boselli, M. Cesarini, F. Mercorio and M. Mezzazanica, Are the Methodologies for Producing Linked Open Data Feasible for Public Administrations?, in: *Proceedings of 3rd International Conference on Data Management Technologies and Applications (KomIS-2014)*, 2014, pp. 399–407. doi:10.5220/0005143303990407.

[10] C. Alexopoulos, L. Spiliotopoulou and Y. Charalabidis, Open data movement in greece: A case study on open government data sources, in: *Proceedings of the 17th Panhellenic Conference on Informatics*, 2013, pp. 279–286–. doi:10.1145/2491845.2491876.

[11] J. Attard, F. Orlandi, S. Scerri and S. Auer, A systematic review of open government data initiatives, *Government Information Quarterly* **32**(4) (2015), 399–418. doi:10.1016/j.giq.2015.07.006.

[12] M. Laessig, B. Jacob and C. AbouZahr, *Opening data for global health*, in: *The Palgrave Handbook of Global Health Data Methods for Policy and Practice*, S. Macfarlane and C. AbouZahr, eds, Palgrave Macmillan, 2019. doi:10.1057/978-1-137-54984-6_23.

[13] M. Jovanovik and D. Trajanov, Consolidating Drug Data on a Global Scale Using Linked Data, *Journal of Biomedical Semantics* **8**(3) (2017). doi:10.1186/s13326-016-0111-z.

[14] H.D.A. dos Santos, M.I.S. Oliveira, G.F.A.B. Lima, K.M. Silva, R.I.V.C.S. Muniz and B.F. Lóscio, Investigations into data published and consumed on the Web: a systematic mapping study, *Journal of the Brazilian Computer Society* **24**(14) (2018). doi:10.1186/s13173-018-0077-z.

[15] A. Varytimou, N. Loutas and V. Peristeras, Towards Linked Open Business Registers: The Application of the Registered Organization Vocabulary in Greece, *International Journal on Semantic Web and Information Systems* **11**(2) (2015), 66–92. doi:10.4018/IJSWIS.2015040103.

[16] A. Hogan, P. Hitzler and K. Janowicz, Linked Dataset Description Papers at the Semantic Web Journal: A Critical Assessment, *Semantic Web Journal* **7**(2) (2016), 105–116–. doi:10.3233/SW-160216.

[17] M. Janssen, Y. Charalabidis and A. Zuiderwijk, Benefits, Adoption Barriers and Myths of Open Data and Open Government, *Information Systems Management* **29**(4) (2012), 258–268.

[18] E. Kalampokis, E. Tambouris and K. Tarabanis, Linked Open Government Data Analytics, in: *Proceedings of the International Conference on Electronic Government (EGOV)*, M.A. Wimmer, M. Janssen and H.J. Scholl, eds, Springer, Berlin, Heidelberg, Koblenz, 2010, pp. 99–110. doi:10.1007/978-3-642-40358-3_9.

[19] A. Zuiderwijk, M. Janssen, S. Choenni, R. Meijer and R.S. Alibaks, Socio-technical impediments of open data, *Electronic Journal of E-Government* **10**(2) (2012), 156–172.

[20] S. Neumaier, J. Umbrich and A. Polleres, Automated quality assessment of metadata across open data portals, *Journal of Data and Information Quality* **8**(1) (2016), 1–29.

[21] B.S. Hitz-Gamper and M.S. O. Neumann, Balancing control, usability and visibility of linked open government data to create public value, *International Journal of Public Sector Management* **32**(5) (2019), 457–472. doi:10.1108/IJPSM-02-2018-0062.

[22] E. Kalampokis, E. Tambouris and K. Tarabanis, On publishing linked government data, in: *Proceedings of the 17th Panhellenic Conference on Informatics (PCI '13)*, Thessaloniki, 2013, pp. 25–32. doi:10.1145/2491845.2491869.

[23] T. Berners-Lee, Linked data, 2006. https://www.w3.org/DesignIssues/LinkedData.html.

[24] C. Bizer, T. Heath and T. T. Berners-Lee, Linked Data - The story so far, *International Journal on Semantic Web and Information Systems* **5**(3) (2009), 1–22. doi:10.4018/jswis.2009081901.

[25] N. Konstantinou and D.E. Spanos, *Deploying Linked Open Data: Methodologies and Software Tools*, in: *Materializing the Web of Linked Data*, C. Springer, ed., 2015. doi:10.1007/978-3-319-16074-0_3.

[26] T. Heath and C. Bizer, *Linked Data: evolving the web into a global data space*, 1st edn, Morgan and Claypool Publishers, Seattle, 2011.

[27] T. Heath, How Will We Interact with the Web of Data?, *IEEE Internet Computing* **12**(5) (2008), 88–91. doi:10.1109/MIC.2008.101.

[28] M. Hausenblas and M. Karnstedt, Understanding Linked Open Data as a Web-Scale Database, in: *Proceedings of the 2nd International Conference on Advances in Databases Knowledge and Data Applications*, Menuires, 2010, pp. 56–61. doi:10.1109/DBKDA.2010.23.

[29] F. Radulovic, M. Poveda-Villalón, D. Vila-Suero, V. Rodríguez-Doncel, R. García-Castro and A. Gómez-Pérez, Guidelines for Linked Data generation and publication: An example in building energy consumption, *Automation in Construction* **57** (2015), 178–187. doi:10.1016/j.autcon.2015.04.002.

[30] W3C, Best Practices for Publishing Linked Data, 2014. http://www.w3.org/TR/ld-bp/.

[31] A.F. Veenstra and T. Broek, *A Community-driven Open Data Lifecycle Model Based on Literature and Practice*, in: *Case Studies in e-Government 2.0*, I. Boughzala, M. Janssen and S. Assar, eds, Springer, Cham, 2014. doi:10.1007/978-3-319-08081-9_11.

[32] AKSW, LOD2: Creating Knowledge out of Interlinked Data, 2014. http://aksw.org/Projects/LOD2.html.

[33] A. Barbosa, I.I. Bittencourt, S.W.M. Siqueira, R.A. Silva and I. Calado, The Use of Software Tools in Linked Data Publication and Consumption: A Systematic Literature Review, *International Journal on Semantic Web and Information Systems* **13**(4) (2017), 68–88. doi:10.4018/IJSWIS.201710010.

[34] V.A. Pinto and F.S. Parreiras, Enterprise linked data: A systematic mapping study, in: *International Conference on Conceptual Modeling*, 2014, pp. 253–262. doi:10.1007/978-3-319-12256-4_27.

[35] J. Jensen, Linked Data in Education: A Survey and a Synthesis of Actual Research and Future Challenges, *IEEE Transactions on Learning Technologies* **11**(3) (2018), 400–412. doi:10.1109/TLT.2017.2787659.

[36] J. Jensen, A systematic literature review of the use of Semantic Web technologies in formal education, *British Journal of Educational Technology* **50**(2) (2019), 505–517. doi:10.1111/bjet.12570.

[37] T.N. Tran, D.K. D. K. Truong, H.H. Hoang and T.M. Le, Linked data mashups: A review on technologies, applications and challenges, in: *Asian Conference on Intelligent Information and Database Systems, ACIIDS 2014*, Springer Verlag, 2014, pp. 253–262. doi:10.1007/978-3-319-05458-2_27.

[38] C. Figueroa, I. Vagliano, O.R. Rocha and M. Morisio, A systematic literature review of linked data-based recommender systems, *Concurrency Computation* **27**(17) (2015), 4659–4684. doi:10.1002/cpe.3449.

[39] A. Zaveri, A. Rula, A. Maurino, R. Pietrobon, J. Lehmann and S. Auer, Quality Assessment for Linked Data: A Survey, *Semantic Web Journal* **7**(1) (2014), 63–93.

[40] K. Petersen, S. Vakkalanka and L. Kuzniarz, Guidelines for conducting systematic mapping studies in software engineering: An update, *Information and Software Technology* **64** (2015), 1–18. doi:10.1016/j.infsof.2015.03.007.

[41] B. Kitchenham and S. Charters, Guidelines for performing systematic literature reviews in software engineering, Technical report, Ver 2.3, EBSE, 2007. doi:10.1145/1134285.1134500.

[42] X. Zhou, Y. Jin, H. Zhang, S. Li and X. Huang, A Map of Threats to Validity of Systematic Literature Reviews in Software Engineering, in: *23rd Asia-Pacific Software Engineering Conference (APSEC)*, 2016, pp. 153–160. doi:10.1109/APSEC.2016.031.

[43] L.C.B. Martins, M.C. Victorino, M. Holanda, G. Ghinea and T.M. Grønli, UnBGOLD: UnB government open linked data: semantic enrichment of open data tool, in: *Proceedings of the 10th International Conference on Management of Digital EcoSystems (MEDES '18)*, K. Peffers, M. Rothenberger and B. Kuechler, eds, ACM, New York, USA, 2018, pp. 1–6. doi:10.1145/3281375.3281394.

[44] R. Fleiner, Linking of Open Government Data, in: *12th International Symposium on Applied Computational Intelligence and Informatics (SACI)*, Timisoara, 2018, pp. 1–5. doi:10.1109/SACI.2018.8441014.

[45] P. Krataithong, M. Buranarach and T.S. T., RDF Dataset Management Framework for Data.go.th, in: *International Conference on Knowledge, Information, and Creativity Support Systems*, 2018. doi:10.1007/978-3-319-70019-9_4.

[46] H. Elmekki, D. Chiadmi and H. Lamharhar, Open Government Data: Problem Assessment of Machine Processability, in: *Information Systems and Technologies to Support Learning. EMENA-ISTL 2018. Smart Innovation, Systems and Technologies*, 2018. doi:10.1007/978-3-030-03577-8_54.

[47] M. Buranarach, P. Krataithong, S. Hinsheranan, S. Ruengittinun and S. Thepchai, A Scalable Framework for Creating Open Government Data Services from Open Government Data

Catalog, in: *Proceedings of the 9th International Conference on Management of Digital EcoSystems (MEDES '17)*, 2017. doi:10.1145/3167020.3167021.

[48] E. Klein, A. Gschwend and A.C. Neuroni, Towards a Linked Data Publishing Methodology, in: *Conference for E-Democracy and Open Government (CeDEM)*, 2016, pp. 188–196. doi:10.1109/CeDEM.2016.12.

[49] A.C.N. Ngomo, S. Auer, J. Lehmann and A. Zaveri, *Introduction to linked data and its lifecycle on the web*, in: *Reasoning Web International Summer School, Lecture Notes in Computer Science, vol 8714*, M.K. et al, ed., Springer, Cham, 2014, pp. 1–99. doi:10.1007/978-3-319-10587-1_1.

[50] S. Sorrentino, S. Bergamaschi, E. Fusari and D. Beneventano, *Semantic Annotation and Publication of Linked Open Data*, in: *Computational Science and Its Applications – ICCSA 2013. Lecture Notes in Computer Science, vol 7975*, B.M. et al., ed., Springer, Berlin, Heidelberg, 2013, pp. 462–474. doi:10.1007/978-3-642-39640-3_34.

[51] M. Kaschesky and L. Selmi, Fusepool R5 linked data framework: concepts, methodologies, and tools for linked data, in: *Proceedings of the 14th Annual International Conference on Digital Government Research (dg.o '13)*, 2013, pp. 156–165. doi:10.1145/2479724.2479748.

[52] H.S. Al-Khalifa, A Lightweight Approach to Semantify Saudi Open Government Data, in: *16th International Conference on Network-Based Information Systems*, 2013, pp. 594–596. doi:10.1109/NBiS.2013.99.

[53] V. Janev, U. Milošević, M. Spasić, S. Vraneš, J. Milojković and B. B. Jireček, Integrating Serbian public data into the LOD cloud, in: *Proceedings of the Fifth Balkan Conference in Informatics (BCI '12)*, 2012, pp. 94–99. doi:10.1145/2371316.2371335.

[54] B. Hyland and D. Wood, *The Joy of Data - A Cookbook for Publishing Linked Government Data on the Web*, in: *Linking Government Data*, B.M. et al., ed., Springer, New York, NY, 2011, pp. 27–49. doi:10.1007/978-1-4614-1767-5_1.

[55] B. Villazón-Terrazas, L.M. Vilches-Blázquez, O. Corcho and A. Gómez-Pérez, *Methodological Guidelines for Publishing Government Linked Data*, in: *Linking Government Data*, B.M. et al., ed., Springer, New York, NY, 2011, pp. 3–26. doi:10.1007/978-1-4614-1767-5_2.

[56] F. Cifuentes-Silva, C. Sifaqui and J.E. Labra-Gayo, Towards an architecture and adoption process for linked data technologies in open government contexts: a case study for the Library of Congress of Chile, in: *7th International Conference on Semantic Systems (I-Semantics '11)*, 2011, pp. 79–86. doi:10.1145/2063518.2063529.

[57] P. Salas, J. Viterbo, K. Breitman and M..A. Casanova, *StdTrip: Promoting the Reuse of Standard Vocabularies in Open Government Data*, in: *Linking Government Data*, B.M. et al., ed., Springer, New York, NY, 2011, pp. 113–133. doi:10.1007/978-1-4614-1767-5_6.

[58] T. Lebo, J.S. Erickson, L. Ding, A. Graves, G.T. Williams, D. DiFranzo, X. Li, J. Michaelis, J.G. Zheng, J. Flores, Z. Shangguan, D.L. McGuinness and J. Hendler, *Producing and Using Linked Open Government Data in the TWC LOGD Portal*, in: *Linking Government Data*, B.M. et al., ed., Springer, New York, NY, 2011, pp. 51–72. doi:10.1007/978-1-4614-1767-5_3.

[59] A.R. Hevner, S.T. March, J. Park and S. Ram, Design science in information systems research, *Management Information Systems Quarterly* **28**(1) (2004), 75–105. doi:10.2307/25148625.

[60] K. Peffers, M. Rothenberger, T. Tuunanen and R. Vaezi, Design Science Research Evaluation, in: *International Conference on Design Science Research in Information Systems*, K. Peffers, M. Rothenberger and B. Kuechler, eds, Springer, Berlin, Heidelberg, 2012, pp. 398–410. doi:10.1007/978-3-642-29863-9_29.

[61] W3C, Data on the Web Best Practices, 2017. https://www.w3.org/TR/dwbp/.

[62] T.M. Yang and Y.J. Wu, Examining the socio-technical determinants influencing government agencies' open data publication: A study in Taiwan, *Government Information Quarterly* **33**(3) (2016), 378–392. doi:10.1016/j.giq.2016.05.003.

[63] K. Forsberg and H. Mooz, The Relationship of System Engineering to the Project Cycle, *Engineering Management Journal* **4**(3) (1992), 36–43. doi:10.1080/10429247.1992.11414684.

[64] T. Berners-Lee, Relational Databases on the Semantic Web, 1998. https://www.w3.org/DesignIssues/RDB-RDF.html.

[65] G. Fu, FCA based ontology development for data integration, *Information Processing & Management* **52**(5) (2016), 765–782. doi:10.1016/j.ipm.2016.02.003.

[66] A. Pivk, Automatic ontology generation from web tabular structures, *AI Communications* **19**(1) (2006), 83–85.

[67] M.C. Pattuelli, A. Provo and H. Thorsen, Ontology Building for Linked Open Data: A Pragmatic Perspective, *Journal of Library Metadata* **15**(3) (2016), 265–294. doi:10.1080/19386389.2015.1099979.