


RESEARCH

Open Access

Investigations into data published and consumed on the Web: a systematic mapping study

Helton Douglas A. dos Santos^{1*†} , Marcelo Iury S. Oliveira^{2,1†}, Glória de Fátima A. B. Lima¹, Karina Moura da Silva¹, Rayelle I. Vera Cruz S. Muniz¹ and Bernadette Farias Lóscio¹

Abstract

The increasing interest in using the Web as a platform for data sharing has motivated research about publishing and consuming data on the Web. While this subject is gaining importance, up until now, there are not many academic papers reviewing the approaches for publishing and consuming data on the Web. Furthermore, to the best of our knowledge, there is no systematic review of the literature that analyzes this subject. In this article, we conduct a systematic mapping study that aims to provide an overview of the current literature on publishing and consuming data on the Web by conducting a systematic mapping study. This study seeks to function as a snapshot of this subject by (i) identifying and analyzing how data have been published and consumed on the Web, (ii) discovering the benefits and limitations of publishing and consuming data on the Web (iii) analyzing the evolution of research on publishing and consuming data on the Web, and (iv) classifying the studies into categories related to their contribution. Finally, we discuss the results of this study and their implications for research on data on the Web-related subjects.

Keywords: Data on the Web, Data consumption, Data publishing, Systematic mapping

Introduction

The World Wide Web has emerged as an important channel for sharing and exchanging information, which has enabled the publication, propagation, and visualization of data from diverse domains [13]. Its rapid growth has been accompanied by the emergence of new paradigms, which seek to ensure that users can take an effective part in making use of the Web [88]. In addition, with the advancement of technology, the data produced by society and made available on the Web has grown rapidly [8]. More recently, the increased publication of Open Data, the large volume of data generated by social networks, the Web of Things (WoT) paradigm, and the Open Web Platform (OWP) paradigm have confirmed the potential of the Web as a platform for sharing and exchanging of data [13, 30, 77].

It is important to note that the interest in publishing and exchanging data on the Web is not new [2, 15]. However, due primarily to the flexibility offered by the Web, new challenges need to be addressed in order to ensure its success as a data sharing platform [13]. The literature contains several studies that set out to investigate issues related with the challenge of publishing data on the Web. Some of these studies propose best practices, guidelines, or processes in order to standardize the data publication process (e.g., [62, 68, 74]). Also, there is considerable research in other issues, such as data cataloging (e.g., [16, 108, 110]), data infrastructure services (e.g., [4, 50, 111]), data integration (e.g., [56]), data linkage and data fusion (e.g., [32, 42]), data publishing (e.g., [97]), and data visualization (e.g., [27, 78]). Moreover, several studies investigate data consumption problems such as data discovery (e.g., [39, 84]), data extraction (e.g., [6, 29, 71]), and data analysis (e.g., [72, 117]). Furthermore, each one of these issues may have multiple research facets.

While the publication and consumption of data on the Web is gaining importance [13], up until now, there

*Correspondence: hdas@cin.ufpe.br

[†]Helton Douglas A. dos Santos and Marcelo Iury S. Oliveira contributed equally to this work.

¹Center for Informatics, Federal University of Pernambuco, Av. Prof. Moraes Rego, 1235 - Cidade Universitária, Recife, PE, Brazil

Full list of author information is available at the end of the article

are not many academic papers that reviewed this subject. In particular, currently available studies focus on some specific approaches such as Linked Data reflecting only a small fragment of the whole set of options related to the publication and consumption of data on the Web [59].

To the extent of our knowledge, Abiteboul et al. [2] were one of the first to use term data on the Web. According to them, data on the Web refers to the use of the Web infrastructure and its set of standards to support data exchange. In particular, they advocate for the use of XML and related standards (e.g., XSLT and XSD) to publish data on the Web. Abiteboul et al. [2] and many other projects conducted at that time led to a significant progress on using Web as a platform to data exchange. In more recent years, the highly development of Web-related technologies opened up various forms of producing, publishing, sharing, and consuming data. In this context, this paper offers a perspective on the contributions on publishing and consuming data on the Web made in the last 11 years. It also describes some of the important bodies of work and outlines' relevant challenges to current data on the Web research. We note in advance that this is not intended to be a comprehensive survey of all the approaches and best practices used to publish and consume data on the Web, and even though the reference list is long, it is by no means complete.

Therefore, in this article, we provide an overview of the current literature on publishing and consuming data on the Web by conducting a systematic mapping study. Systematic mapping is a protocol-driven methodology for reviewing and synthesizing a research data area [67]. A systematic mapping study typically provides an overview of the research reported in the field and identifies possible issues arising from examining the existing literature. This study seeks to function as a snapshot of the data on the Web publication and consumption by (i) identifying and analyzing how data have been published and consumed on the Web, (ii) discovering the benefits and limitations of publishing and consuming data on the Web (iii) analyzing the evolution of research on publishing and consuming data on the Web, and (iv) classifying the studies into categories related to their contribution.

The rest of the work is organized as follows. In the “[Theoretical background](#)” section, we discuss the theoretical background. In the “[Related works](#)” section, we present the related works. In the “[Research approach](#)” section, we describe the research methodology used in our study. In the “[Results](#)” section, we present the result analysis. A discussion about open issues is presented in the “[Discussion and research directions](#)” section. Finally, in the “[Conclusions](#)” section, we present our conclusions.

Theoretical background

In general, articles on publishing and consuming data on the Web often refer to a set of best practices and approaches closely aligned to the general architecture of the Web [63, 75]. In summary, Jacobs and Walsh [63] state that the Web is composed of a set of resources uniquely identified by Uniform Resource Identifiers (URIs), whose representation can be usually retrieved via standardized formats. A resource representation encodes information about the resource, and state is usually typed. HTML, RDF, XML, or CSV are examples of data formats. Agents or users can interact with Web resources using standardized protocols, which control the exchange of messages HTTP, FTP, and SOAP, for example. Messages include both data and metadata. The development and use of such standards enable the Web to transcend different technical architectures. It is possible to use generic data browsers to explore the data available on the Web [18], for example. In fact, both humans and machines can gather data from Web.

In particular, data published on the Web deals with specific types of Information Resources¹ called datasets, which are collections of data, published or curated by a single agent, and available for access or download in one or more formats [76]. A dataset does not have to be available as a downloadable file, but it can be accessed through a Web API or data stream. Moreover, data should also be available in machine-readable formats, provided in a convenient form, and offered without technological barriers for data consumers. Therefore, data must be released in formats that reasonably structure the data, but that also allows automated processing and facilitates machine sorting and searching activities.

In the following, we present relevant aspects related to publishing and consuming of data on the Web.

Data on the Web lifecycle

Within the Web environment, there are several activities that make up the process of publishing and consuming data, ranging from dataset planning and creation to access and process of datasets. According to Lóscio et al. [73], the set of these activities is called the lifecycle of data on the Web, during which data are being created, published, exported, imported, consumed, processed, and reused by different parties and for different purposes. In this way, understanding the lifecycle allows a better understanding about the nature of the data as well as provide a shared vocabulary that allows different practitioners to discuss about essential issues related to the publishing and consumption of data on the Web. In addition, a data lifecycle helps to explain paradigm shifts, to compare the functionality of different platforms, and to aid the integration of previously disparate implementation efforts [82].

Möller [82] proposes the Abstract Data Lifecycle Model (ADLM), which is a generic model for lifecycle representation for data and metadata, establishing a common set of phases, characteristics, and roles. According to ADLM, a lifecycle for data-centric domains must consist of the ontology development, planning, creation, archiving, refinement, publication, access, external use, feedback, and termination phases. Due to its generic nature, it can be used to construct new data-centric lifecycle models.

However, some ADLM phases are not suitable for all data environments. According to Lóscio et al. [73], the Web Data context does not include ontology development, archiving, and termination phases. Moreover, Lóscio et al. [73] state that the ontology development is an independent activity and therefore was not included. In a similar way, the archiving and termination phases were not considered because once the data has been published on the Web, it should be always available. In this sense, Fig. 1 presents a lifecycle for data published and consumed on the Web proposed by Lóscio et al. [73] and based on ADLM model. Although it is a cycle, it is possible that not all steps are followed until a new iteration begins. Thus, even though it is a cycle, this does not mean that the data have to go through the last stage before starting a new iteration or that feedback needs to be received just before the producer refines the data. During this process, actors play the role of data producer and consumer, where, in general, the producer is responsible for creating or publishing the data. On the other hand, consumers are responsible for consuming the data, and may also be producers, since they can make improvements to and refine the data in order to publish them again [73].

All phases are briefly described below.

- *Planning*: Ranges from the intention to publish the data to the selection of the data that will be published [73]. Lóscio et al. [73] points out that it is important to take into account the potential of data usage and, where possible, to ask potential data consumers to identify relevant data.
- *Creation*: Ranges from data extraction phase to data transformation (i.e., transforming data into appropriate format for Web publishing). The creation phase also comprises the metadata creation, which will describe the data [73]. It is important to consider publishing in different formats (distributions), minimizing the need for the transformation of data by consumers [74].
- *Publication*: Makes data available on the Web in a form for (re)use by others. It involves the tasks focused on keeping the data accessible. Often, data cataloging tools are used to publish data [73]. To guarantee the appropriate access to metadata, it is advised to provide a suitable search engine to retrieve these data. It also may involve controlling the access to data.
- *Access*: Consists of the act when users gain access to the data [73].
- *Consumption*: Comprises series of actions and methods related to the manipulation and analyses of the collected data. In fact, it is the actual use of the data. This stage of the lifecycle is directly related to the data consumer. Among consumers, we can mention from a developer interested in creating an application that makes use of the data, as people interested in transforming the data to generate relevant information, as well as large companies interested in using data to improve their products

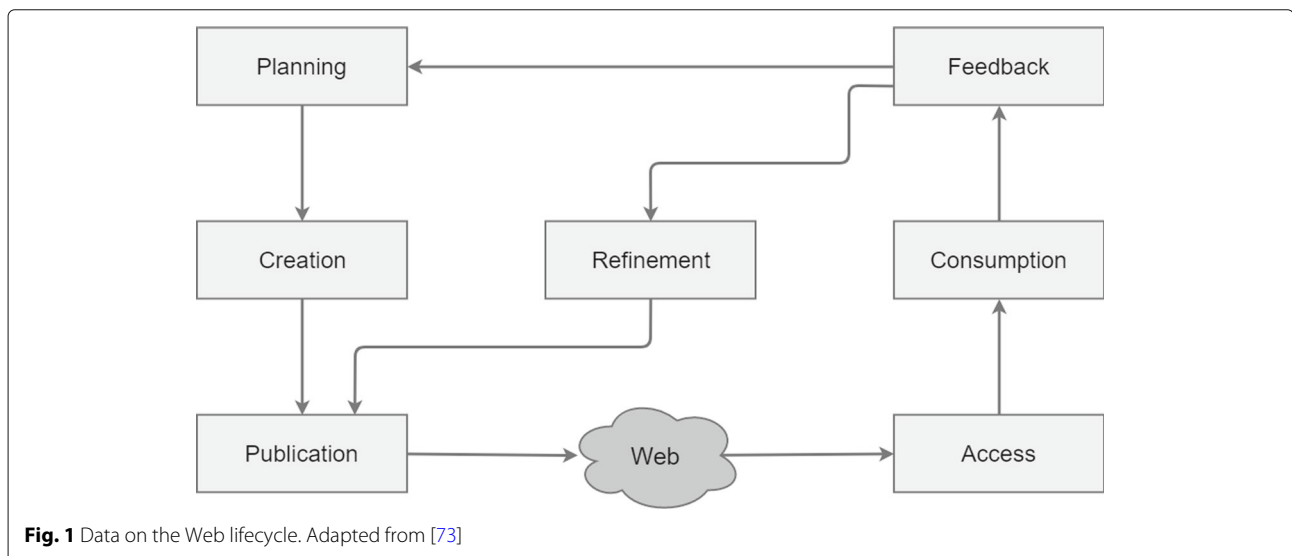


Fig. 1 Data on the Web lifecycle. Adapted from [73]

and services and, even another system that consumes the data.

- *Feedback*: Comprises the moment when consumers should provide comments on the data and metadata used, allowing to identify improvements and corrections in the published data, as well as to maintain a channel of communication between producers and consumers [73].
- *Refinement*: Comprises all activities related to improvements and updates into published data. It is related to the guarantee of maintenance of the previously published data and that can be realized from the comments of *feedback* collected in the previous phase. In addition, Lóscio et al. [73] state that refinement can be done either by generating new versions to ensure that the data is not obsolete, by correctly managing the different versions or by providing access to the correct version of the data for consumers.

Data on the Web Ecosystem

The data on the Web Ecosystem may be defined as a set of actors and artifacts involved in producing, distributing, and consuming data by using the Web [75]. An actor can be a user, a system, or a device and can act either as a data producer or as a data consumer. The former delivers and produces data of some type according to specific conditions. The latter consumes (e.g., processes, analyzes, filters, aggregates) data. Both actors interact with each other by exchanging datasets.

Examples of data on the Web ecosystems are Open Government Data Initiatives [8]. In these initiatives, governments act as data producer making their data available in machine-readable formats and under open licensing conditions that allow the use and redistribution of the data. Entrepreneurs, application developers, or citizens act as data consumers by creating new information products and services, visualizations, and mash-ups that allow to monitor the activities of their government, as well as creating tools to make daily life easier. Such available government data have the ability to facilitate networks of collaboration and co-creation to produce citizen empowerment as well as promote accountability and other democratic principles.

Producers are responsible for data publication activities, such as defining licenses, choosing formats, and platforms for distribution. Furthermore, they can provide one or more access interfaces to retrieve data. Each interface determines the requirements to be satisfied by data consumers in order to successfully use the service, such as parameters, outputs, and operations. Moreover, data consumers consume data according to specific requirements, which are the conditions and the capabilities needed to solve a problem or achieve an objective.

In order to consume or produce datasets, consumers and producers, respectively, must coordinate a set of activities, which represent a piece of work that forms one logical step (i.e., operation) within a consumption or production process. According to Dittrich and Jonscher [41], choosing a particular set of activities may depend on the intended use of the data, the capability of the actor, the characteristics of the data (e.g., alphanumeric data, multimedia data; structured, semi-structured, unstructured data), requirements concerning data quality, service performance requirements, and the available resources (e.g., human resources, time, money)

In summary, data on the Web Ecosystems rely on a vast and heterogeneous set of actors, each one with different properties, capabilities, and expectations. Similarly, datasets are heterogeneous regarding structural (schema), syntactic (format), and semantic (meaning) issues. Actors may produce and consume a dataset using different activities and under different conditions. Also, many of these elements are dynamic and evolve with time. We may conclude that the data on the Web Ecosystem landscape is one of the distributed, heterogeneous, dynamic, and evolving actors and resources.

The emergence of data on the Web Ecosystems has been driven by several factors, including the emergence of digital technologies and political/institutional initiatives. For instance, the majority of Data Ecosystems have been driven mostly by the open data movement, which calls for free use, reuse, and redistribution of data by anyone [55]. Several governments already launched Open Data Portals to stimulate and promote Open Data production and consumption [33]. The technology improvement (e.g., mobile Internet or technology) and technology trends (e.g., social media or mobile apps) also have been driving private and public organizations to publish data as well as to integrate their services with external data.

Related works

There are some studies that analyze the state of the art of some of the approaches used to publish and consume data on the Web. For instance, Bizer [18] presents an overview on the major Linked Data providers and also the few efforts on how to build applications that exploit the Web of Linked Data. In [19], the authors extended [18] analysis by presenting conceptual and technical principles of Linked Data. They also situate these principles within the broader context of related technological developments. Moreover, they also reviewed applications developed to exploit and publish Linked Data. More recently, Bikakis and Sellis [17] describe the major pre-requisites and challenges that should be addressed by up-to-date approaches to exploring and visualizing very large linked datasets.

There are some studies that review specific application domains on Linked Data. For instance, Barnaghi et al. [12]

review some of the recent developments on applying the Linked Data technologies to the Internet of Things. In particular, they focus on the information modeling, ontology design, and processing of semantic data problems. They describe the initial progress and some of the developments. They also discuss the future prospects and challenges of developing efficient semantic-enabled IoT systems. Similarly, Bröring et al. [25] illustrates and analyzes the recent developments on applying Linked Data principles for sensor technologies. Bröring et al. [25] also point out challenges and future topics for research on semantic sensors.

From the Open Data perspective, Zuiderwijk et al. [119] present a survey study on the impediments to using Open Data. According to them, the main Open Data issues are (1) availability and access, (2) discoverability, (3) usability, (4) understandability, (5) quality, (6) linking and combining data, (7) comparability and compatibility, (8) metadata, (9) interaction with the data provider, and (10) opening and uploading. Geiger and Von Lucke [51] also look at the general challenges of Open Data publication. With regard to social aspects, Conradie and Choenni [35] focus on the understanding of how internal processes influence data publishing. They found that data publishing costs are still associated in terms of locating data and, in some cases, getting permission to publish data which might prohibit the release of large amounts of data. From the data consumption perspective, Zuiderwijk et al. [118] provide an overview of the barriers that data consumers (*citizen*) may encounter in using public sector information, such as lack of knowledge of the data, or no knowledge about its existence. Zhang et al. [115] also report various thresholds to the release of data, including the lack of tools for sharing and conflicting data definitions.

As to Open Government Data, Petychakis et al. [94] present a state-of-the-art analysis of open government data infrastructure from a functional, semantic, and technical perspective. The authors focus on the current open government data landscape of European Union countries. Another example is the study [24], which presents a survey of existing Open Government Data platforms, focusing on the technical aspects. The authors took into account features such as standardization, discoverability, and machine readability. According to them, most platforms lack of proper standards and Application Programming Interfaces (APIs), and there is a significant amount of data published either using non-machine-readable format or in a proprietary format [24].

Janssen et al. [64] analyze the Open Data benefits as well as the main barriers faced by Open Data initiatives. They synthesized user experiences with Open Data obtained from interviews and a group session. A large number of benefits of Open Data was identified. In particular, political and social benefits were viewed as the most important

category. With regard to barriers, the complexity of handling the data, the use of Open Data and participation in the Open Data process, the legislation and the quality of information were identified as the main problems. Janssen et al. [64] also analyzed the myths of Open Government Data, especially the myth that publicizing data will automatically yield benefits.

These studies focus on analyzing individual projects, applications, or conceptual ideas and specific contexts such as Linked Data and Open Data. There are other approaches to publish data on the Web, such as Big Data [13, 20] and Web Semantic Sensor Data [107], and there are specific communities that publish and consume data on the Web without following Linked Data or Open Data principles, such as research data communities [9]. A systematic and holistic review is necessary in order to provide further insights on the current state of research related to publish data on the Web as well as the overall development of the topic, which can form the basis for shaping future works. To the best of our knowledge, until now, there has been no systematic literature review on data on the Web, in order to map the state of the art, alongside the identification of research gaps and expected benefits.

Research approach

The scientific literature differentiates at least two types of systematic reviews: conventional systematic reviews and mapping studies. Conventional systematic reviews, aggregate results about the effectiveness of a treatment, intervention, or technology and are related to specific research questions (e.g., *Is intervention "I" on population "P" more effective for obtaining outcome "O" in context "C" than comparison treatment "T"?*) [93]. Mapping studies aim to identify all research related to a specific topic, i.e., to answer broader questions related to research trends [7]. Typical questions are exploratory, (e.g., *What do we know about topic "T"?*).

In this paper, we performed a mapping study with the aim of identifying the scenario in which data on the Web have been published and consumed over the last 11 years as well as of discovering problems, barriers and obstacles faced when publishing and consuming data on the Web.

Research Questions

We used the following Research Question (RQ) to guide our processes for searching and selecting studies: RQ: *What has been the scenario of publishing and consuming data on the Web over the last 16 years?*

We then used seven specific research questions to guide and structure data extraction, analysis, and the synthesis of all the evidence:

- RQ1: How has publishing and consuming data on the Web research evolved over the last years?

- RQ2: What are the main types of contributions reported by the studies?
- RQ3: What are the characteristics of data published and consumed on the Web?
- RQ4: What methods or procedures have been used for publishing and consuming data on the Web?
- RQ5: What tools have been used for publishing and consuming data on the Web?
- RQ6: What is currently known about the barriers and limitations related to publishing and consuming data on the Web?
- RQ7: What are the main benefits related to publishing and consuming data on the Web?

Inclusion and exclusion criteria

From an initial set of 8740 papers, we selected studies presenting concepts, theories, guidelines, discussions, lessons learned, and experience reports on publishing and consuming data on the Web (inclusion criteria). We excluded papers that fell into any of the following criteria:

- 1 Paper is not written in English;
- 2 Paper cannot be accessed on the Web;
- 3 Paper was not published between 2005 and 2016;
- 4 Paper is not peer-reviewed work (e.g., invited papers, keynote speeches, workshop reports, books, theses, and dissertations);
- 5 Paper is incomplete documents, drafts, slides of presentations, and extended abstracts;
- 6 Paper addresses other areas besides Computer Science (e.g., social science, health-care, and others);
- 7 Papers that do not present any type of findings on publishing and consuming data on the Web.

Data sources and search strategy

The search process combined automatic and manual search to achieve high coverage. The manual search was conducted on journals and conferences (see Table 1 for a complete list of sources considered in our manual search). In particular, manual searches are important to cover the cases where published papers are available in the manual sources but have not yet been indexed by the search engines used in the automatic search. We looked for titles, abstracts, and keywords of all papers in each source used in the manual search, using the same procedure applied to the list of papers returned by the automatic search. The use of manual search is supported in the literature on systematic reviews so as to complement and extend the coverage of automatic searches [67, 93]. The automatic search was performed in five search engines and indexing systems (see Table 2 for a complete list sources considered in our automatic search).

The search string (see Fig. 2) used in the automatic search consists of the phrase construction “Data on the

Table 1 Manual sources

Sources
<i>Journal of Organizational Computing and Electronic Commerce</i>
<i>Journal of Information Technology and Politics</i>
<i>Governance. An International Journal of Policy, Administration and Institutions</i>
<i>Transforming Government: People, Process and Policy</i>

Web” and terms related to contemporary approaches used to publish and consume data on the Web (i.e., Open Data, Linked Data, Open Government Data). In particular, these terms were inspired by the classification proposed by Lóscio et al. [75], which states that Data on the Web can be viewed as a most broad set of data published according to the Web architecture [75]. Figure 3 illustrates the relationship between Data on the Web, Open Data, and Linked Data. As we may observe, not all data available on the Web is shared openly nor follow Linked Data principles. In other words, data publishers determine the policy upon which data will be shared.

We also used the predicate, which consists of synonyms of “publication” and “consumption,” and we used wildcard characters to capture the plural and singular forms of the keywords. The query string is intentionally kept simple so that we can extract the maximum number of papers containing the terms. The search string was adapted for each search engine.

We constructed the search string using several iterations and pilot tests to ensure that we used a comprehensive set of synonyms to allow for high coverage while keeping the number of retrieved articles under control. Considering the high number of results from our automatic search process (over 8740 articles), we believe that we achieved a reasonable coverage level with our search string in the automatic search.

Selecting studies

Six researchers performed the manual and automatic searches working in partnership on a given engine or set of manual sources. As automatic sources have features for filtering papers according a date interval, we applied the third exclusion criterion, thereby excluding studies not published between 2005 and 2016. Note that

Table 2 Automatic sources

Sources	URL
IEEE Xplore	https://ieeexplore.ieee.org/Xplore/home.jsp
ACM Digital Library	https://dl.acm.org/
Springer	https://link.springer.com/
ScienceDirect	https://www.sciencedirect.com/
Scopus	https://www.scopus.com/

("Data on the Web" OR "Open Data" OR "Linked Data" OR "Open Gov Data" OR "Open Government Data" OR "Web of Data") AND ("Publication" OR "Consumption" OR "Publishing" OR "Consuming" OR "Data Publisher" OR "Data Provider" OR "Data Consumer" OR "Data Producer")

Fig. 2 Search string. Source: Author

in automatic search, we preferred to opt for a wider search string, a task we gave to the researchers, with a view to reducing the chance of losing relevant articles. Many of these articles were eliminated in the primary selection, where the researchers evaluated the results from the automatic search ($n = 8740$) and from the manual search ($n = 194$) by looking at the title and abstract and excluding the papers that were either clearly not relevant or duplicated ones. This resulted in 193 relevant studies.

The list of 193 potentially relevant studies was analyzed for final selection. All six researchers worked on the selection process. Initially, the 193 papers were divided into subsets of 65, and each subset was assigned to one pair of researchers, there being 3 pairs. In each pair, each researcher worked independently on analyzing the papers in their assigned set. The researchers applied the inclusion and exclusion criteria (see the "Inclusion and exclusion criteria" section) on the potentially relevant papers after reading the abstract, introduction, and conclusion of each paper. Differences were solved during a consensus meeting. This process selected 46 papers considered relevant for data extraction and analysis (the whole study selection process is described in Fig. 4).

Data extraction and data synthesis

Data extraction was carried out guided by an extraction form implemented in MS Excel™. In this step, the 6 researchers were also divided into the previously defined three pairs to independently analyze each paper in order to answer the research questions previously defined. Conflicts arising from extracting information were discussed and solved in consensus meetings. The results from data extraction were analyzed with support of MS Excel™, which was also used to generate graphics.

In the synthesis of results, we used mind maps, constructed from the data extracted and made available on the spreadsheet, in order to synthesize and obtain a better understanding of these data. During this process, the extraction spreadsheet was divided into 6 parts. Each researcher was responsible for synthesizing one part. The parts were coded with respect to the topics or themes addressed. Categories that emerged were combined using constant comparison techniques.

Threats to validity

The most common threats in a systematic mapping are the coverage of the study, possible research biases in the study selection process, and inaccuracies during the phases of data extraction, analysis, and synthesis. These were also

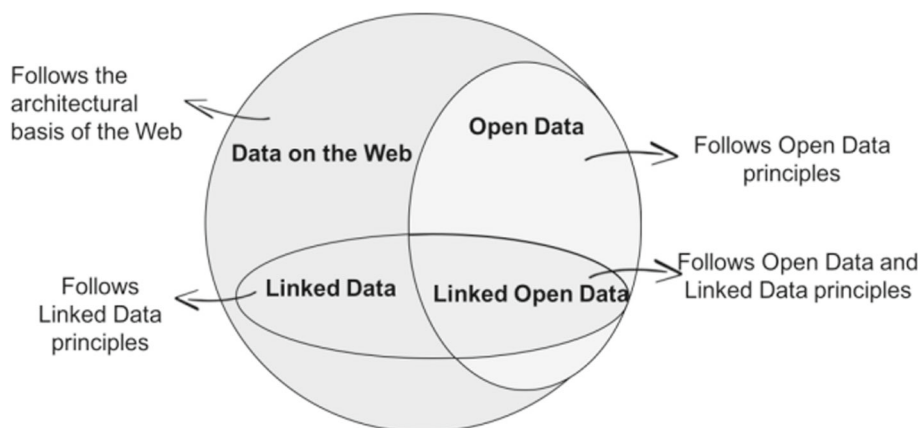


Fig. 3 Data on the Web × Open Data × Linked Data. Source:[75]

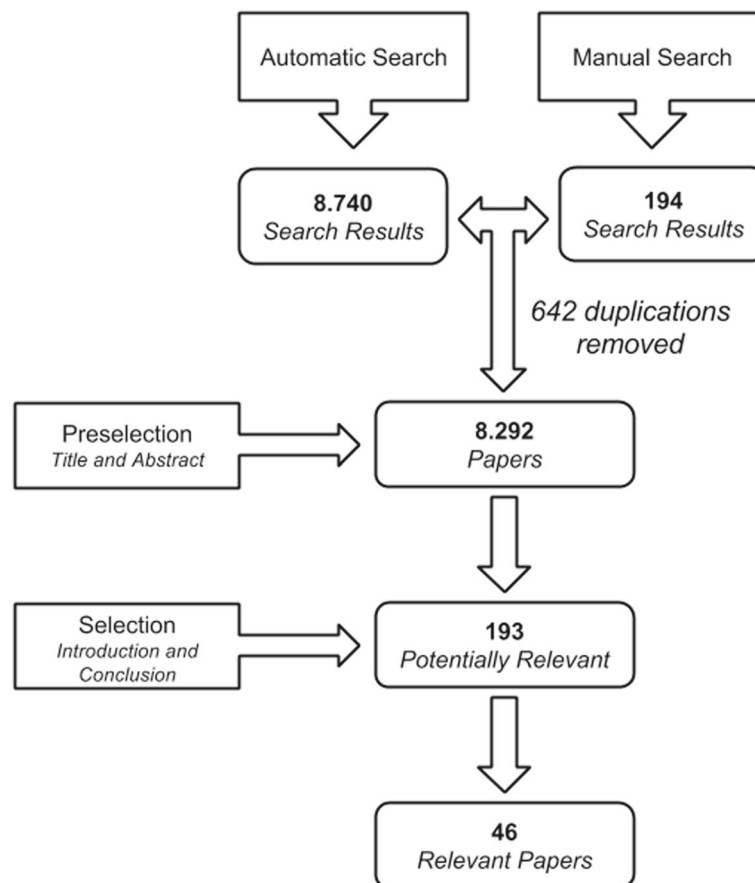


Fig. 4 Study selection. Source: Author

the main threats in our study. We tried to minimize them by using well-established guidelines for our research protocol.

The combination of automatic search in the most popular search engines, and manual search in relevant publication venues, improved mapping coverage. However, coverage is an inherent threat to validity in any systematic review or mapping. In general, it is not possible to reach 100% of coverage, and it is very difficult to estimate reliably the coverage within a literature review.

The data extraction process was carried out by two researchers, and the conflicts were solved by third parties or in consensus meetings. Since the reviewed studies use different terminology, the extraction of data is more prone to errors, especially when analyzing the data extracted in order to answer RQ1. In particular, there are limitations related to using methodologies, techniques or tools, due to the fact that some papers do not mention or describe which were used.

The query string is intentionally kept simple, with broad terms and simple predicate, so we can extract the maximum number of papers containing the terms. However, it is possible that some paper had been eliminated in the pre-selection phase. This problem may have occurred because of the title and abstract of paper do not present a clear mention of Web as the platform for sharing data. In addition, some articles may have focused only on discussing a specific problem related to the data process itself (i.e., data integration) rather than addressing particularities of Web as data sharing platform. On the other hand, there are also studies that report the improvements on their solutions or databases by using Linked Data principles; however, the derived data are not published on the Web. Studies of this type were also eliminated in the selection phase.

Results

This mapping review analyzed 46 research papers, published between 2005 and 2016. In the following sections, we present the main results of our study. We discuss each one of the research questions presented in the

“**Research Questions**” section based on the selected studies presented in Table 3.

How has data on the Web research evolved over recent years?

This section reports both on the descriptive information, temporal, and geographic distribution of papers and on the methodological issues arising from reviewing the primary studies.

We analyzed the evolution of publications over the years (see Fig. 5). [W46] is the first data on the Web study, considering the range of years from 2005 to 2016, and was published by Rajiv C. Shah, Jay P. Kesan, and Andrew Kennis in 2008. They analyze how a city in Massachusetts treated Open Standards for document formats. They developed a number of lessons so that other governments can consider introducing a similar policy. [W7] is the last selected work which presents a framework called LinDA that aims to hide the underlying complexity of Linked Data while maintaining and promoting the interlinking capabilities enabled by the Linked Data paradigm. This framework allows an ecosystem of Linked Data for the Public Sector to be created. Such ecosystems provide

significant benefits to data consumers, such as increased accessibility and reuse of data, Web-scale identifiers and easy interlinking with datasets of other producers of public data.

We consider that having a total number of 46 articles on such an important research subject is very small. However, the number of studies might increase because we did not consider studies that have been published in 2017. Moreover, with the exception of the decrease in 2011, the average number of publications lightly fluctuates around some mean over the years.

Which publication venues are the main targets?

Table 4 presents the venues that were used for publishing the studies more than once. Over 67.39% (31 studies) of the papers were published in journals. Fifteen papers were published in conferences and workshops (see Table 3). Forty-one studies are full papers, and four are short papers.

With the exception of *Datenbank-Spektrum*, *International Journal on Digital Libraries*, *Journal of Biomedical Semantics*, and *Records Management Journal* and the

Table 3 Summary of selected papers

ID	Reference	Year	Venue	ID	Reference	Year	Venue
W1	[116]	2010	Journal	W24	[60]	2011	Conference
W2	[58]	2010	Journal	W25	[109]	2012	Conference
W3	[43]	2013	Journal	W26	[65]	2013	Conference
W4	[23]	2012	Journal	W27	[95]	2014	Conference
W5	[21]	2012	Journal	W28	[26]	2014	Conference
W6	[22]	2015	Journal	W29	[85]	2014	Conference
W7	[86]	2016	Journal	W30	[79]	2014	Conference
W8	[98]	2015	Journal	W31	[34]	2014	Conference
W9	[66]	2010	Journal	W32	[103]	2014	Conference
W10	[37]	2015	Journal	W33	[31]	2014	Journal
W11	[112]	2013	Journal	W34	[101]	2014	Journal
W12	[113]	2013	Journal	W35	[69]	2015	Journal
W13	[49]	2015	Journal	W36	[81]	2015	Journal
W14	[53]	2016	Journal	W37	[36]	2015	Journal
W15	[3]	2015	Journal	W38	[99]	2015	Journal
W16	[92]	2015	Journal	W39	[114]	2016	Conference
W17	[10]	2016	Journal	W40	[38]	2016	Conference
W18	[61]	2016	Journal	W41	[5]	2014	Conference
W19	[47]	2014	Journal	W42	[44]	2014	Journal
W20	[54]	2015	Conference	W43	[83]	2014	Journal
W21	[28]	2013	Journal	W44	[48]	2015	Journal
W22	[14]	2010	Conference	W45	[100]	2016	Conference
W23	[104]	2013	Journal	W46	[105]	2008	Journal

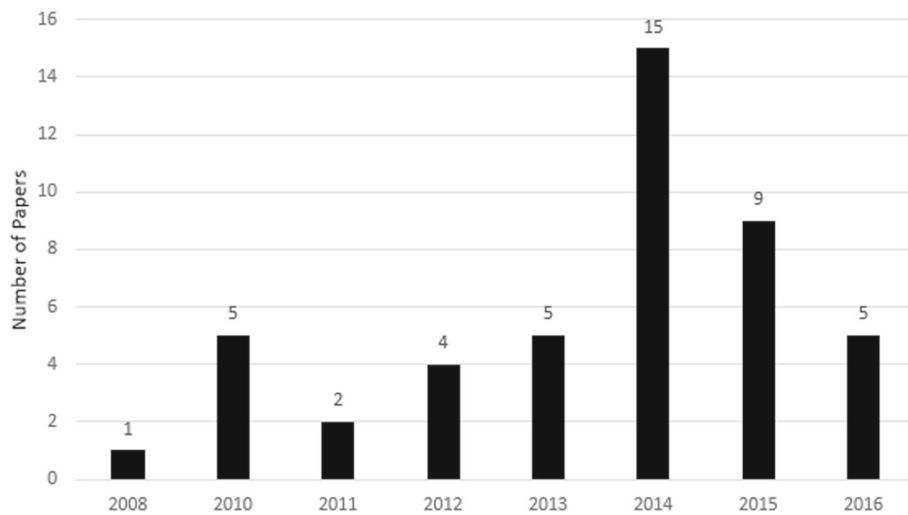


Fig. 5 Temporal distribution of works over the years. Source: Author

International Conference on Semantic Web: Trends and Challenges that were used as a venue for publishing two studies each, the other venues published only one paper each.

What individuals, organizations, and countries are most active in data on the Web research topic?

In the 46 papers reporting data on the Web, 171 distinct co-authors were identified. The most active researchers are Asunción Gómez-Pérez, Daniel Vila-Suero, Lisa Raymond, María Poveda-Villalón, Spiros Mouzakitis, and Varsha K. Khodiyar each of whom co-authored 2 studies, while all other authors co-authored only 1 study. The authors belonged to 93 distinct organizations (universities, research institutions, and companies) located in 28 different countries. The Faculty Getulio Vargas, National Technical University of Athens, The Campus, Universidad Politécnica de Madrid, and University of Southampton were the most active organizations with 2 studies each, with the exception of the Universidad Politécnica de Madrid which is associated with 4 studies. Finally, the 6 most active countries were China and Japan (2 studies each), Greece, Italy and Netherlands

(3 studies each), Spain (4 studies), Germany (6 studies), USA (8 studies), and the UK (10 studies).

What are the main types of contributions reported by the studies?

Regarding research question RQ4, we are interested in examining what kind of results the studies are reporting. They were classified based on a taxonomy adapted from [80, 106], which describes the kind of contribution a study makes. Each study can present one or more contributions, which means that it may or may not contain features of other contribution types. Hence, the types of contribution were divided into the following categories:

- Analysis: A primary study that analyzes an object of study as regards a structure or taxonomy, a method, a framework, or any set of evaluation criteria;
- Comparative study: A primary study that compares static analysis approaches to identify cases in which the application of one approach is better than another;
- Improve existing tool: A primary study that aims to improve an existing tool;
- Method: A primary study that searches for a general solution for a problem area, such as a process, guidelines, best practices, maturity model, methodology or well-grounded checklists;
- New tool: A primary study that proposes a new tool created or implemented by applying some method or technique, which may be more effective than existing tools or may used in combination with other tools;
- Report: A primary study documenting knowledge and experience obtained, rules of thumb or checklists.

Table 4 Venues used more than once for the publications

Venue type	Venue name	Works
Conference	International Conference on Semantic Web:	
	Trends and Challenges	W29, W31
Journal	<i>Datenbank-Spektrum</i>	W2, W4
Journal	<i>International Journal on Digital Libraries</i>	W8, W17
Journal	<i>Journal of Biomedical Semantics</i>	W12, W21
Journal	<i>Records Management Journal</i>	W33, W42

Table 5 relates the contribution of categories and studies reviewed. The data collected indicate that most contributions are analysis (18 papers), the proposal of methods (14 papers) or tools (13 papers), and reporting of studies (10 papers). As to the papers that propose a method, the large majority focus on the data publication problem (12 studies). In particular, 11 papers propose a process, guidelines, best practices, or some kind of engineering methods and techniques that are to be used to manage a data publication or consumption problem. In addition, 13 studies propose a tool to assist data on The Web practitioners. However, while 6 of the studies present a concrete solution like a platform or framework, 4 remain at the conceptual level as they present a reference architecture. Moreover, 2 papers focus on proposing an improvement of an existing tool. In fact, these papers describe recent developments in adapting an existing solution to address Linked Data principles.

Table 5 Contributions of the research studies

Contribution category	Works
<i>Analysis studies</i>	
Barriers, challenges, and common issues	W5, W6, W10, W13, W17, W22, W34, W39
Data model	W9
Data on the Web cases	W3, W15, W20
Landscape	W2, W3, W7, W23, W26, W33, W37
Method	W45
<i>Comparative studies</i>	
Data legal and political frameworks	W16
Methodologies for publication of data	W35
<i>Improvement of existing tool</i>	
Addressing linked data features	W11, W12
<i>Method</i>	
Data model	W6, W9, W21
Data on the Web engineering process	W1, W17, W18, W19, W21, W25, W28, W30, W31, W35, W38, W45
<i>New tool proposal</i>	
Architecture	W8, W26, W36, W43
Data consumption tool	W4, W21, W29
Platforms and frameworks	W14, W22, W24, W36, W40, W44
<i>Reporting works</i>	
Linked data experiences	W1, W6, W22, W25, W27, W40, W41, W46
Open data experiences	W32, W42

The analysis studies focus either on a specific object of study (e.g., the publishing of research data on bio-informatics) or on the landscape about a data on The Web approach. Moreover, 16 of them analyze data publication problems, namely the focus is on the data consumption process (1 study) and 1 study analyzes both publication and consumption issues. In particular, the analysis papers focus on study of the landscape (7 papers) or benefits and barriers about a data on the Web approach (8 papers), data on the Web case (2 papers), on the analysis of a specific method (1 paper) or data model (1 papers), and on the study of best practices (4 papers). There are also studies that report data on the Web initiatives. Seven papers document Linked Data experiences, and 2 papers report Open Data initiatives. Finally, 1 study presents comparative documentation of the Open Data legal and political frameworks of Argentina and Brazil, and 1 study makes a comparative study of the existing methodologies and the best practices for publishing open government data.

Overall, the types of contributions of most of the studies are weak: lessons learned, tools, and guidelines as well as examples of overview analysis such as their benefits, limitations, and landscape. Moreover, there are 11 papers that focus on documenting a data on the Web initiative/experience. Of the remaining studies, only 16 studies exhibit a stronger contribution type, such as a theory, a framework/method or a model.

What are the characteristics of data published and consumed on the Web?

Most of the data published on the Web belong to specific domains. When analyzing the studies, we observed that the vast majority belong to the governmental and academic domain, as shown in Table 6. As a result of the increase in the number of open government data initiatives, the number of publications involving government data is booming. In all, we have identified 22 papers dealing with government-related data.

The academic domain is producing and making data available. It also consumes data every day. Of the articles selected, 21 of them cover data from the academic

Table 6 Data domains

Domains	Papers
Academy	W1, W2, W3, W4, W5, W6, W9, W10, W11, W13, W14, W17, W18, W19, W22, W29, W30, W31, W32, W33, W45
Government	W4, W6, W7, W15, W16, W21, W24, W25, W26, W27, W28, W32, W34, W35, W36, W39, W40, W41, W42, W43, W44, W46
Institution	W8, W9
Without specific domain	W12, W20, W23, W37, W38

world. An example is paper [W40] which shows an analysis of Linked Data practices in education. Another domain found in two papers was the institutional one. Paper [W8] shows a Linked Data architecture proposal to be used in the archives of the Getulio Vargas Foundation. Some papers (W4, W6, W9, and W32) covered more than one domain, others did not determine any particular domain, as described in Table 6.

In addition to the data domains, we sought to find out which data publishing approach to made data available on the Web. We identified 12 papers deal only with Open Data, while 8 are dedicated to Linked Data. However, most articles used the Linked Open Data hybrid approach, totaling 21 papers (see Table 7).

We identified five papers (W14, W17, W18, W30, and W46) that publish data on the Web, but are not related with Linked Data or Open Data. As a matter of fact, with the exception of W14, the rest of these five papers focus on research data publishing. According to Murphy et al. [87], “Research data publishing is the release of research data, associated metadata, accompanying documentation, and software code [...] for reuse and analysis in such a manner that they can be discovered on the Web and referred to in a unique and persistent way.”

Paper [W14] proposes a collaborative and open database, called OpenTrials, to publish structured data and documents on all clinical trials. Paper [W17] analyzes the current data publishing workflow landscape across disciplines and institutions in order to present the generic components of such workflows, i.e., to provide a reference model for these stakeholders. Despite advocate for neither Open Data nor Linked Data Principles, the authors recommend the use of existing standards for repositories and all parts of the data publishing process, and the development of new standards where necessary. In its turn, [W18] proposes and show examples of changes to the format and peer-review process for journal articles to more robustly link them to data that are only available on request. However, despite data items are closed, the dataset must be public indexed in the Web. Paper [W18] also proposes additional features for data repositories to

better accommodate non-public clinical datasets. Paper [W30] discusses the facilities approach to managing and publishing research data. Paper [W30] uses the notion of Research Objects, which can be defined as resources that bring together data, methods, and people in scientific investigations.

Paper [W46] chronicles the historic process of Massachusetts Government to adopt open standards to format dataset, spreadsheets, charts, presentations, and word processing documents. However, this paper do not employ the open principles for access, use, or reuse of data.

In addition, after analyzing the papers, we verified that the Open Data approach was more used in the papers with a governmental domain, while linked data and linked open data were applied more in the academic domain. Figure 6 shows the distribution of papers per approach.

An analysis was also performed on what data formats are used to publish and consume data. Data formats are ways to make data available for consumption. There are formats of various types, i.e., those with structured, unstructured, and semi-structured data. Some of these formats are proprietary, such as Microsoft Excel file format (XLS). In this analysis, several types of formats were mentioned, but the Resource Description Framework (RDF) format excelled in areas that favor structured and connected data. In the Open Data approach, Comma Separated Value (CSV) is one of the most cited publication formats, as we can see in Fig. 6. There are studies that reference more than one data format. For instance, the papers [W6][W31] employ CSV, JavaScript Object Notation (JSON), and RDF formats. Moreover, 13 studies do not specify explicitly a specific data format.

The use of vocabularies when publishing data on the Web is a frequent practice, especially in the context of Linked Data. Vocabularies are important for structuring the data and describing its domains. Each of these vocabularies is described by a document that has a URI for each defined resource. In this analysis, 43 different vocabularies were cited, such as FOAF¹, Data Catalog Vocabulary (DCAT)², Bibliographic Ontology (BIBO)³, Simple Knowledge Organization System (SKOS)⁴, Vocabulary of Interlinked Datasets (VOID)⁵, and Geonames⁶. The most cited in the articles was FOAF, a vocabulary aimed at describing people and their personal relationships, with a total of 6 articles that mentioned it. However, other vocabularies have also been highlighted, such as SKOS cited in 5 articles, VOID was cited in 4 articles, and, finally, DCAT was cited in 4 articles (see Table 8).

As to data consumption, we analyze the way data is made available and accessed. The most common approach identified was the use of SPARQL Protocol and RDF Query Language (SPARQL) endpoints. The popularity of this approach is due to the fact that many of the

Table 7 Data publishing approaches

Areas covered	Papers
Linked data	W5, W11, W20, W22, W23, W27, W28, W43
Linked open data	W1, W2, W3, W6, W7, W8, W12, W13, W21, W24, W25, W26, W29, W31, W36, W37, W38, W40, W41, W44, W45
Open data	W4, W9, W10, W15, W16, W19, W32, W33, W34, W35, W39, W42
Other data publishing approaches	W14, W17, W18, W30, W46

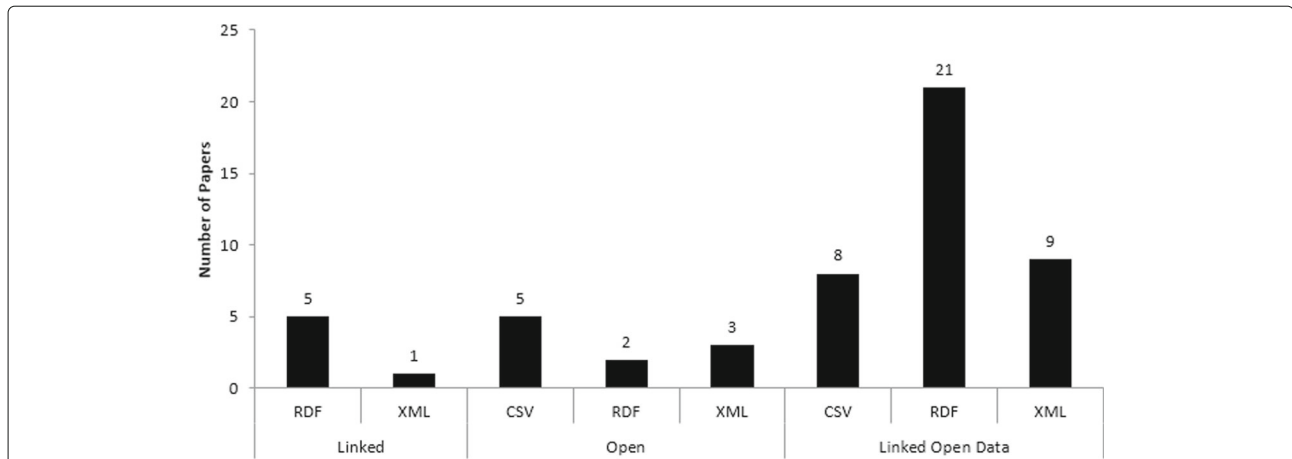


Fig. 6 Publishing formats by area. Source: Author

papers analyzed use Linked Data or Linked Open Data approaches. Of the papers reviewed, 18 of them (36.7%) cited SPARQL as a reference language. However, as mentioned earlier, many of the papers use the CSV, JSON, etc. formats and consequently make these files available for direct download.

What methods or procedures have been used for publishing and consuming data on the Web?

Methodologies for publishing data on the Web are generally used to guide processes for publishing data, as well as provide guidelines and best practices, thus making the craft of the data publisher viable. When analyzing the studies, we found that more than half of the papers, that is, 28 of them (61%) do not mention the use of data publication methodologies in the Web in their studies. In Table 9, we can verify that 18 papers (39%) cited the use of methodologies, while 14 used Guidelines, 2 cited or used Best practices, and 4 papers described general methodologies. Some works reference more than one methodology, such as [W1] and [W37].

Of the 14 papers that followed guidelines in their studies, 12 are related to Linked Data principles, created by Berners-Lee, outlining a set of “rules” where Linked Data

publishers must adopt them in the publication process. This shows that a good many of the studies related to publication of Linked Data are applying and following Linked Data principles.

On the other hand, 2 studies cited or made use of Best practices. Paper [W11] describing its research method states that it applied Best practices described by the W3C Health Care and Life Sciences Interest Group in its methodological process. However, paper [W35] is a survey of methodologies and Best practices for publishing open data. Thus, it only cites the existence of Best practices for Publishing Linked Data, which it describes as a methodology widely used in Linked Data and Linked Open Data.

Finally, 4 papers cited the use of general methodologies, that is, they were not classified as guidelines or Best practices. Among these papers, paper [W37] stands out because it makes a comparative study between methodologies applied in Linked Open Data and Digital Libraries and thus references 12 methodologies, in which 5 guidelines are included.

What tools have been used for publishing and consuming data on the Web?

Both in publishing and in data consumption on the Web, tools help publishers and data consumers to exercise their roles more quickly and productively. In all, 18 articles cited the use of these tools, some of which mentioned

Table 8 Most cited vocabularies

Vocabularies	Papers
BIBO	W21, W40
DCAT	W31, W36, W38, W45
FOAF	W8, W21, W26, W44, W40, W45
GeoNames	W26
PROV	W8, W41
SKOS	W3, W8, W26, W41, W44
VoID	W1, W25, W31, W36, W38, W40

Table 9 Methods and procedures used in the studies

Methods and procedures	Papers
Best practices	W1, W2, W3, W8, W21, W24, W25, W26, W27, W28, W37, W40, W44, W45
Guidelines	W11, W35
Others	W1, W15, W31, W37

more than one tool in their study. Table 10 presents the five most cited tools among the 18 papers that clearly used data publishing and consumption tools on the Web.

Apache Jenna¹, the open source Java framework for Semantic Web, provides features that make it possible to read and create RDF triples from other formats. Thus, in the context of the lifecycle of data on the Web [73], Apache Jenna covers the creation phase, as it enables data to be created in the appropriate format for publication. In addition, it also reaches the consumption phase, as it makes it possible to read the data.

The D2R Server² is a framework that allows relational databases to be consulted using Semantic Web standards. It rewrites requests in SQL queries. This tool considers, according to the lifecycle of data on the Web [73], the consumption phase, since it allows the consumer to make queries in different ways.

Pubby³ provides Linked Data interfaces for SPARQL endpoints. Thus, like the D2R Server, it also acts in the consumption phase of the data published on the Web, thereby providing better visualizations for Linked Data.

Silk⁴ is an open source framework for integrating heterogeneous data sources. By using this tool, data publishers can establish RDF links from their data sources to other data sources on the Web. Therefore, like Apache Jenna, it also belongs to the data creation phase.

SPARQL Endpoint allows its users to run SPARQL queries on datasets in the Linked Data format. In general, it is used to conduct more accurate queries, since the user must be proficient in the SPARQL language. Many Web data providers provide this endpoint to conduct searches on their data. Finally, these tools are part of the consumption phase of the data lifecycle on the Web.

What is currently known about the barriers and limitations related to publishing and consuming data on the Web?

Due to the technological expansion and growth of the Web of data, new challenges are emerging, such as making data available so that consumers can easily find, access, and use them. Among the 46 papers analyzed, 12 of them did not cite or comment on challenges faced by the community. In our analysis, we found 34 challenges that were cited and reported in the studies, which divided them into problems and barriers. In Table 11, we can see the 5 most

Table 10 Most widely used tools

Tools	Papers
Apache Jena	W25, W41
D2R server	W3, W26
Pubby	W1, W25
Silk	W1, W25, W26
SPARQL endpoint	W2, W3, W29

Table 11 Most cited challenges of publishing and consuming data

Challenges	Papers
Data quality problems	W2, W22, W33, W37, W42
Data sharing problems	W8, W9, W17, W18, W33
Interoperability problems	W2, W3, W5, W8, W13, W14, W16, W17, W22, W30, W34, W36, W37
Political and social barriers	W7, W17, W26, W34, W35
Technical barriers	W4, W7, W16, W32, W34, W35

cited challenges in the papers. The interoperability was a problem that was most highlighted in our analysis, as it obtained 13 citations from the 34 papers.

Interoperability between datasets is a problem that is especially common when we talk about connected data, but it is also found in other areas. In our analysis, 13 papers cited and described problems with interoperability (see Table 11). Among the problems, paper [W5] describes there being a large number of entities which have unique access URIs. In addition, the study reports that different RDF graphs are constructed using different vocabularies, which also reflects on a semantic problem of the data. Paper [W3] adds that securing the data format and the compliance of its structure to heterogeneous sources is a major challenge within the Linked Data universe, as is providing the integration of this data with multiple providers [W2] [W17] [W22]. As a solution to this whole context of problems related to interoperability between connected data, paper [W36] suggests that the data provider must create a common architecture that can be applied to other data platforms, thus allowing the data to be linked. Out of the context of connected data, paper [W16] shows that interoperability issues also exist on accessing open data on the Web.

Technical barriers are easily found when publishing and consuming data on the Web and were cited by 6 papers (see Table 11). Paper [W16] describes this barrier as a negative parameter of open data access initiatives on the Web. The great effort needed to adopt tools, and their complexity is a barrier embedded in Linked Data tools, paper [W7] reports. Non-skilled users or consumers also suffer from these barriers. Paper [W4] notes that there is currently no single system that supports non-specialized users such as journalists to search for and analyze heterogeneous and distributed public datasets. Within the Web, there are datasets that have machine-readable formats and others do not, and this results in some difficulties for the reuse of data, according to paper [W34]. This paper also points out that this is a very common and easy to find technical barrier. Leaving to the public context, paper [W32] states

that the public sector has great difficulty in providing useful sites based on their data, thus having a negative impact on their open access initiatives. In addition, paper [W35] shows us that there are issues and challenges related to the technology used, the data formats adopted, and the infrastructure needed for publishing open government data.

The political and social barriers faced by publishers and consumers of data on the Web were cited by 5 papers. Paper [W7] describes that the mindset in the public sector needs to change due to the bureaucracy as well as to the complex legislation that characterizes public services. It also points out that public agencies are prone to adopting new technologies, such as Linked Data, at a much slower pace than private organizations and companies do. Paper [W34] says that there are challenges regarding accessibility and the reusability of public sector information. Paper [W35] has given a more general description and reported that there are indeed challenges related to political support, as well as to decision-making and to social problems.

In our analysis, 5 papers cited problems with quality of data published on the Web. When we talk about data quality, we are talking about both the dataset and the data itself. It is common to come across challenges or problems with data quality when consuming or searching for published data on the Web, whether open or not. For paper [W33], the quality of data in the context of open data publishing is still a challenge. In addition, data quality is undoubtedly a key point in publishing datasets of any Open Data project [W42]. The paper also shows that this problem can only be solved when information is created and that managing data quality after the process of creation and publication is a rather complex task. Paper [W22] points to the lack of data or incompleteness of the data as indicative of the poor quality of the dataset as well as of correctly entered data. As one of the solutions to problems with data quality, paper [W2] mentions that merging data can help resolve data conflicts as well as quality problems. Paper [W37] reports that the lack of metadata is one of the factors that cause data to be of poor data quality and that is important to have metadata to aid in the management and sharing of published data.

Another challenge cited by 5 papers is the act of sharing data through the Web. According to paper [W9], data sharing is a complex and theoretically challenging objective. Papers [W8], [W18], and [W33] also deal with data sharing and reuse as a new challenge faced by the scientific community. In addition, paper [W17] reports that repositories and journals are generally not ready to deal with sensitive data, and it is important that the data sharing mechanism is appropriate to the sensitivity level of such data.

What are the main benefits related to publishing and consuming data on the Web?

In recent years with the growth of data publication on the Web, the benefits for users and publishers have stood out and encouraged stakeholders to exercise new roles on the Web. In this systematic mapping, 28 benefits were identified, such as discovering services, better data quality, better feedback, improvements in research, efficiency, greater trust, economic benefits, better public services, sharing, an increase in reproducibility, accountability, visibility, the development of applications, ease in aggregating data, collaboration, accessibility, reuse, transparency, innovation, openness, easy discovery, and interoperability. Table 12 shows the 6 benefits most cited in the 46 articles, which are discussed below in detail.

Being able to discover data easily is the most cited benefit. The process of data discovery relies on searches made by end users and intermediate actors in order to acquire knowledge or develop new solutions. Nowadays, even though the volume of data is increasing exponentially, it is becoming easier to find data on multiple sources on the Web and filter them to one's purpose. The use of good practices and the growth of Web Semantic have helped data discovery in machine-readable formats, gathered and interpreted automatically [W37], and this has also helped it to discover datasets more easily from different domains, such as government bodies, public companies, private organizations, and in health and scientific matters [W17] [W27]. This benefit collaborates with other benefits, such as better visibility, reuse, and sharing.

Re-using data on the Web is the second benefit most cited in studies. The reuse of data enables the recycling of data that were previously published, aggregating and adding value to them. Reuse of data is a solution for the enormous number of datasets shared on the Web that get lost without being used. This benefit enables new sources to be created by integrating data with other data [W13],

Table 12 Most cited benefits of publishing and consuming data

Benefits	Papers
Better reuse of data	W4, W6, W7, W9, W13, W14, W15, W16, W23, W24, W26, W27, W30, W31, W32, W33, W37, W38, W41, W42, W44, W45
Easy discovery	W1, W2, W3, W9, W11, W13, W14, W15, W17, W18, W19, W20, W21, W23, W24, W26, W27, W29, W30, W36, W37, W38, W41, W44, W45
Economic benefits	W1, W4, W5, W8, W10, W13, W14, W15, W23, W27, W32, W33, W35, W36, W39, W40
Interoperability	W6, W7, W8, W11, W13, W16, W17, W20, W21, W26, W27, W33, W34, W36, W37, W45
Social benefits	W3, W4, W5, W6, W8, W9, W10, W13, W14, W15, W16, W17, W18, W19, W20, W21, W27, W30, W33, W34, W35, W36, W37, W39, W40, W41, W44, W42, W45

it facilitates research [W14] and disseminates Semantic Web paradigms [W37] and [W38]. Furthermore, this benefit is reached optimally when legal and technical restrictions are removed, enabling new forms of collaboration and innovation, thereby increasing the value of data whenever it is reused and linked to another source [W15].

Social benefits are the third most cited benefits. Social benefits are a group of benefits that improves the interaction between data handlers and the data itself, through practices and data attributes. Social benefits are data sharing, promoting openness, enhanced collaboration [W9], increased visibility, greater visibility [W18], and greater transparency [W13]. These benefits improve the way data consumers, publishers, and intermediaries handle and deal with the data and their products and services. The scientific community, funding agencies, governments, and society are being benefited by these social benefits and increasingly cite them, thus adding value to products and services, roles in science, reproduce research, make data accessible, and advance and accelerate research studies and innovation [W13]. Society's trust increases when governments are more transparent [W10]. When the data become available, success in using them depends on the social and professional context of its community of users [W9] and the audit tools available for the community [W16] in order to help their decision-making [W42]. Participation and collaboration from the community help to promote Open Data, governmental transparency and create innovation through public usages [W39]. Study [W6] shows that the effectiveness of aid programs can be improved by providing transparent insight into aid activities.

Interoperability occurs when data systems are able to work together in a transparent way. Interoperability helps to promote transparency, add value, and make data available [W8] and promotes data sharing and the development of new applications. Interoperability is improved when concepts of the Semantic Web are applied to data on the Web, such as when schemes and vocabularies map concepts and relations [W6] and [W37], and standards are adopted at both the machine-to-machine and data levels [W13]. Also, this benefit enhances when the focus from the systems managing the data is switched onto the data themselves [W37] and the data publishing process follows best practices to help to maintain and sustain data over time [W17].

The economic benefits have grown over the years due to new solutions, with products and services and data being treated as a valuable good. Since the open data movement started, many data domains that interest society, organizations, and agencies are being published and are generating new ways of developing products and services. Information extracted from datasets published on the Web are sources of economic indicators for governments,

companies, and society in order to show economic trends and failures that otherwise go unnoticed, and they promote transparency. Also, the spread of economic information influences and educates the community, transforming the way the citizens stands up to government and making it more accountable for its policies and actions [W4]. Innovation, accountability, new business, and new investments are examples of economic benefits generated by data on the Web [W5], [W13], and [W23].

Discussion and research directions

The main goal of this study is to provide an overview of how data are being published or shared on the Web. To do so, a systematic mapping was performed in accordance with the search method described in the “[Research approach](#)” section. In addition to addressing relevant research questions that have allowed us to delineate an overview of the area, we have also identified some topics that are not addressed in the body of the literature, where many questions were not answered and require further research. In this section, we discuss the results of this study and their implications for research on data on the Web.

Theoretically, within our research and analysis, we found papers in the literature that report experiences that have occurred within the lifecycle of data on the Web, whether addressing one or more phases. We note that there is a greater interest in exhibiting experiences of publishing data, in different formats, sources, and domains. Few studies took the precaution of detailing the whole process and thus left aside the methods, tools, and procedures they used. Many studies used the Linked Data standards, thereby justifying the high index of using the RDF format (see the “[What are the characteristics of data published and consumed on the Web?](#)” section). In this context, we are faced with several problems related to publishing and consuming data. Among these, the search for better interoperability is currently a crucial point in the research conducted on data on the Web, since the discovery and sharing of data in a practical and efficient way is the key to fostering the culture of reuse (“[What is currently known about the barriers and limitations related to publishing and consuming data on the Web?](#)” section). On the other hand, innumerable benefits brought about both in the publication and consumption phases were identified in the papers analyzed (among which ease of access and discovery of data were the benefits with the highest number of citations).

Regarding to the lifecycle of the data on the Web, we could see that few papers focused their contribution on the phases of access, consumption, feedback, and refinement of data. In many cases, their focus included only the phases of planning, creating, and publishing data. As a result, we identified some aspects that are barely

approached in the literature, which include the absence of solutions for monitoring the consumption of the published data, maturity and governance models, managing data on the Web, data and metadata curation, and Data Ecosystems.

Among the aspects identified, we found that there is a need for additional research and that such research should also consider the lifecycle presented, which thus underpin their studies. In this context, there are important areas that should be considered in the development and evolution of publishing and consuming data on the Web: Data Ecosystem theory and models, Data on the Web Management System, Monitoring the consumption of data, Metadata Curation, Maturity Models, and Collaborative refinement of data on the Web. In the following, we present a discussion about these areas.

Data Ecosystem theory and models

The data lifecycle is closely related to the creation of a Data Ecosystem, which relies on a vast and heterogeneous set of actors (e.g., data consumers and data producers) and resources (e.g., datasets, software, and services), each of which has different properties, quality, and functional requirements (as mentioned in the “[Data on the Web Ecosystem](#)” section, Data Ecosystems). Being able to effectively organize and categorize the Data Ecosystem will ultimately deliver more intelligence to industry, academy, and governments [89, 90]. For instance, traditional retailers, telecoms, banks, and other companies are tailoring their services and products based on knowledge and facts extracted from several data collections available on the Web [70]. Governments are using Open Data collections to promote democratic principles such as transparency, accountability, and responsiveness [50].

However, despite the fact that Data Ecosystems are thus arguably gaining in importance, research on Data Ecosystems is still in its preliminary stages. Up until now, not many academic papers related to the Data Ecosystem field have been published. In most cases, they are focused on some component technology that reflects only a small fragment of the whole research area. The same is as true for Data Ecosystems Theory that should provide a conceptual basis for further field research. The terminology and definitions for Data Ecosystem vary greatly. This diversity poses a pressing problem for the development of a clear understanding about how to exploit the new opportunities and emergent challenges in Data Ecosystems. Accurate definitions are required in order to get a mutual understanding of what Data Ecosystems embody.

Moreover, designing, developing, and further maintaining systems for supporting Data Ecosystems is challenging. For instance, no one can responsibly consume data without accompanying information that explains how the

data have been created, where it is located, details about the structure and meaning of the data, and how to collect, integrate, and analyze the data. A comprehensive and meaningful description of all actors and resources is needed. Models help in understanding the functioning and activities of Data Ecosystems. They also help practitioners reduce misunderstandings and have a kind of blueprint for running and managing Data Ecosystems. Data Ecosystem models should support the practitioners and researchers in having a proper idea about the current state of a Data Ecosystem. They may also help to define strategic planning towards achieving the goals of an ecosystem, such as value creation and new businesses. Hence, a model tends to provide the means for developing a framework to control and manage an ecosystem.

Data on the Web Management System

Generally speaking, having a system for managing data on the Web is a somewhat bold idea, but a much-needed one, since the lifecycle of data is being affected by bad publishing practices. While there are Best practices for publishing data on the Web [74], many producers still do not have knowledge of them nor they give the attention they should to these practices, which may well lead to future consequences that mainly impact their consumers.

A data on the Web Management System would make it easy to define, create, maintain, manipulate, and share datasets on the Web across multiple users and applications. As an alternative for implementing one, it should consist of a collection of services that allow users to share datasets on the Web [91]. There are research studies that target this area, but nothing concrete. Some propose separate services, both for data manipulation and for sharing. Future research could unite all these services into a single system that is able to perform the whole set of lifecycle activities. There are also activities for which an architectural model have not yet been proposed and have not been implemented.

Monitoring the consumption of data

Tracking the use of datasets and applications using this data is still a big challenge. Such information could be very useful for the identification of new datasets as well as for the data quality improvement. Monitoring data consumption, as well as providing effective ways for the consumer to interact with the data publisher, should make it possible to collect information about using and sharing data. In this sense, it is crucial to obtain consumers' feedback in such a structured way that allows identifying supposed flaws in the published data, the need to publish new data and to enable classification of data, for example.

After collecting information, it is possible to obtain resources that guide how to refine the monitored dataset,

thereby ensuring that the data are maintained and corrected. In the analysis of the papers, we did not find solutions that covered this aspect. Some data catalog solutions (e.g., CKAN and Socrata) provide only simple options for the registration of comments or submitting contact forms as a means for consumers to send feedback comments. We did not find studies proposing solutions to collect structured and accurate feedback from data consumer nor proposing alternatives to leverage such feedback to (semi)automatically improve the published data. Therefore, we now list some questions which can guide future research studies in this area as follows: (i) What strategies should be defined to perform this monitoring? (ii) What are the main benefits of monitoring the use of data on the Web?

Metadata curation

Heterogeneity, scale, and dynamicity issues make it difficult or, in some cases, hinder the sustainability of Data Ecosystems. Metadata have been seen as critical to the continued success of Data Ecosystem initiatives [40, 102]. The evolving and complex nature of a Data Ecosystem raises the need for an infrastructure to support the management of metadata. Metadata is the foundation for harnessing the vast and diverse amounts of data before they become unmanageable [40, 102]. When metadata are available, the objects (e.g., actors and datasets) that they describe can be rapidly located and accessed for new applications, for instance. This is why metadata need both to be preserved and to be readily available over a long period of time in order to be properly used by the Data Ecosystem participants. Moreover, when metadata are efficiently managed and preserved, this enables them to be discovered and properly reused.

A promising solution is to use a well-conceived, efficient curation strategy for metadata. Metadata curation is the continuous process of managing, improving, and enhancing the metadata and their use [1, 46]. Furthermore, the metadata curation process aims to ensure that the metadata meet a defined set of quality requirements, such as security rules, integrity constraints, or metadata availability expectations. Without proper curation, metadata may deteriorate in terms of their quality and integrity over time. One of the major challenges towards achieving efficient and continuous curation of metadata is to create a methodology to structure the curation process as well as to provide a set of tools to support the curation process.

A metadata curation process needs to give those who use and contribute to the metadata a sense of ownership and control [52]. By contributions is meant the authority to capture and collect metadata, to preserve them, to conduct analysis arising from using metadata everyday, and other tasks related to maintaining metadata. The

systematic curation of metadata requires both an underlying metadata model to describe all the aspects related to the underlying Data Ecosystem and a curator-friendly and efficient environment. Furthermore, the ongoing increase in size and complexity of the Data Ecosystem is forcing its actors towards a common environment in which to manage metadata. Hence, a metadata curation environment that offers tools and guidance should help Data Ecosystem actors with a shared understanding of the available metadata to take responsibility for maintaining metadata in the long term.

Maturity models

Maturity models are widely used in organizational and software engineering. According to [45], maturity models are used to compare and evaluate improvements, thus allowing the degree of evolution in certain domains to be measured. In the business environment, they aim to help organizations identify ways to improve the quality of their processes and reduce their execution time, thereby providing them with competitive advantages.

In general, since data quality is a crucial factor for data consumption and sharing, it is important that there is a prior assessment not only of the data, but of the dataset as a whole. In this context, maturity models can be applied to provide clear recommendations on how to drive improvements based on knowledge of the maturity level in which the dataset lies.

Maturity models can also be applied to assess maturity of data on the Web ecosystems through the use of metrics and levels. These models should follow a thematic approach, for example, they should measure the degree of management and coordination of an ecosystem, from simple ad hoc coordination to formally defined management processes. Thus, they can all be applied to provide guidance to drive improvements based on the knowledge of the maturity level in which the ecosystem is located.

Maturity models are also used in the context of Big Data published on the Web to aid its deployment, as well as offering insights. For example, TDWI [57] is a Big Data Maturity Model that aims to provide all the structure necessary for organizations to understand where they are, where they have been, and what they still need to accomplish in Big Data.

However, in this literature review, no papers were found that cited or used maturity models with the objective of solving quality problems related to the publishing and consuming data on the Web. We also did not find any maturity model for Data Ecosystem. In this context, there is a need for research directions in these areas, since innumerable benefits can be gained by using maturity and governance models within the Web.

Collaborative refinement of data on the Web

Refinement is one of the 7 phases of the lifecycle of data on the Web. It is directly related to maintaining the published data, as well as updates and adding new data. According to [73], this refinement can be carried out based on the feedback provided by the data consumer to the publisher. On the other hand, according to [96], in order to have an ecosystem of data on the Web in fact, there must be an interaction cycle, which involves consumers and publishers in sharing reusable data, with a view to refining them.

Given these two views of refinement, we can highlight the importance of a collaborative environment within the data on the Web ecosystem. Datasets that often have anomalies in their original source can be cleaned and refined as a result of this collaboration between publishers and consumers, both through feedback, and by refining datasets collaboratively.

The aim of collaborative refinement is to give the data consumer the role of refiner as well. Therefore, when data consumers have a problematic set of data at hand, they need to make the necessary adjustments and to correct anomalies before the dataset may be used for other purposes. Finally, after this process, the consumer may request a new publication of the dataset, which now no longer has the problems from its original source.

However, after analyzing the papers collected, we found little content on the area of refining data on the Web, much less collaborative refinement. On the other hand, we found papers regarding the reuse of data as a benefit that often leads to the data refinement. In this scenario, related research is needed to better define and understand this collaborative approach with the emphasis on providing more efficient Data Ecosystem.

Conclusions

The way that individuals and organizations are producing, sharing, and consuming data on the Web has changed with the advent of new technologies. As a consequence, data have come to be seen as a significant and valuable good. This study is a report on a systematic mapping of literature of work on publishing and consuming data on the Web. We found and analyzed 46 relevant studies from a gross total of 8292 extracted from a list of online library databases and conference proceedings.

The purpose of this work was to provide an overview of the field and identify possible research issues or areas not covered. To achieve this goal, we analyzed the evolution of research on this area and also classified the papers according to their research contribution and how they addressed relevant topics. We reported a quasi-steady number in published research from 2014 to date, i.e., with the exception of the decrease in 2011, the average number of publications lightly fluctuates around some mean

over the years. We also described the data on the Web domains and listed the benefits and barriers reported in the literature.

There are some gaps in the literature that may hinder the consumption and production of data on the Web. This study reveals difficulties that should be considered in order to facilitate how data on the Web are published and consumed, including data quality and processes of consumption and publication of data. Even after years of publishing data on the Web and the influence of the Semantic Web movement, data quality is still a challenge. Several studies (e.g., [11, 91]) identified that a significant set of the data published by some data sharing initiatives are available using unstructured, proprietary, and non-machine-readable formats. Another problem is the lack of studies about maturity models that could improve data quality on publishing and consuming data on the Web. There is also lack of exploratory works describing data consumption and usage experiences. Such studies could provide important information about benefits and limitations on this matter. Finally, to the best of our knowledge, there is a lack of widely followed standards and guidelines to publish data on the Web and a lack of use of solutions published in previous studies on similar areas or on the area of data on the Web.

Endnotes

¹ Information Resources are identifiable resources whose essential characteristics can be conveyed in a message [63].

¹ <http://xmlns.com/foaf/spec/>

² <https://www.w3.org/TR/vocab-dcat/>

³ <https://lov.okfn.org/dataset/lov/vocabs/bibo>

⁴ <https://www.w3.org/TR/skos-primer/>

⁵ <https://www.w3.org/TR/void/>

⁶ <http://www.geonames.org/ontology>

¹ <https://jena.apache.org/>

² <https://d2rq.org/d2r-server>

³ <http://wifo5-03.informatik.uni-mannheim.de/pubby/>

⁴ <https://silkframework.org/>

Abbreviations

APIs: Application programming interfaces; BIBO: Bibliographic ontology; CSV: Comma separated value; DCAT: Data catalog vocabulary; FOAF: Friend of a friend vocabulary; HTTP: Hypertext transfer protocol; JSON: JavaScript object notation; NoSQL: Not only SQL; OWP: Open web platform; RDF: Resource description framework; RQ: Research question; SKOS: Simple knowledge organization system; SPARQL: SPARQL protocol and RDF query language; URI: Uniform resource identifier; VOID: Vocabulary of interlinked datasets; WoT: Web of things; XLS: Microsoft excel file format

Acknowledgements

This research study was partially supported by CAPES, CNPq, FACEPE, and INES. Helton Santos receives a master fellowship granted by CNPq. Karina

Moura receives a master fellowship granted by CAPES. Glória de Fátima receives a master fellowship granted by CNPq. Rayele Vera receives a master fellowship granted by FACEPE. Marcelo Iury receives a doctorate fellowship granted by CNPq and CAPES. The authors would also like to thank the colleagues of the Aladin research group for their input for this paper.

Funding

This work was partially supported by funds from the following Brazilian funding agencies: the Pernambuco State Science and Technology Support Foundation (FACEPE), the National Institute of Science and Technology for Software Engineering (INES), Committee for Technological and Scientific Development (CNPq), and Coordination for the Improvement of Higher Level Personnel (CAPES). These funding agencies supported the research by granting doctorate and master fellowships to the majority of authors.

Availability of data and materials

All the electronic documents, spreadsheets, and mindmaps used for coordinate and the systematic mapping activities will be publicly available at <https://github.com/marceloiury/dataontheweb>.

Authors' contributions

Professor BF supervised all the research. HD and MI coordinated the research activities, were the primary writers and participated in all research activities. The rest of the authors were involved with the major part of the activities related to the systematic mapping ranging from selection to the analysis of papers. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details

¹Center for Informatics, Federal University of Pernambuco, Av. Prof. Moraes Rego, 1235 - Cidade Universitária, Recife, PE, Brazil. ²Academic Unit of Serra Talhada, Federal Rural University of Pernambuco, Rua Dom Manuel de Medeiros, s/n, Dois Irmãos, 24105 Recife, PE, Brazil.

Received: 16 November 2017 Accepted: 24 September 2018

Published online: 07 November 2018

References

- Abbott D (2013) What is digital curation. <http://www.dcc.ac.uk/resources/briefing-papers/introduction-curation/what-digital-curation/>. Accessed 10 Jan 2018
- Abiteboul S, Buneman P, Suciu D (2000) Data on the Web: from relations to semistructured data and XML. Morgan Kaufmann Publishers Inc., San Francisco
- Alexopoulos C, Loukis E, Mouzakis S, Petychakis M, Charalabidis Y (2015) Analysing the characteristics of open government data sources in Greece. *J Knowl Econ* 9:1–33. <https://doi.org/10.1007/s13132-015-0298-8>
- Amazon (2015) Amazon web services. <http://aws.amazon.com/es/ec2/>. Accessed 10 Jan 2018
- Aracri R, De Francisci S, Pagano A, Scannapieco M, Tosco L, Valentino L (2014) Publishing the 15th Italian population and housing census in linked open data. In: Proceedings of 2nd International Workshop on Semantic Statistics, Riva del Garda
- Arasu A, Garcia-Molina H (2003) Extracting structured data from web pages. In: Proceedings of the 2003 ACM SIGMOD international conference on Management of data. ACM, San Diego. pp 337–348. <https://doi.org/10.1145/872757.872799>
- Arksey H, O'Malley L (2005) Scoping studies: towards a methodological framework. *Int J Soc Res Methodol* 8(1):19–32
- Attard J, Orlandi F, Scerri S, Auer S (2015) A systematic review of open government data initiatives. *Gov Inf Q* 32(4):399–418
- Austin CC, Brown S, Humphrey C, Leahey A, Webster P, Fong N (2015) Research data repositories: review of current features, gap analysis, and recommendations for minimum requirements. *IAASSIST Q* 39(4):24–38
- Austin CC, Bloom T, Dallmeier-Tiessen S, Khodiyar VK, Murphy F, Nurnberger A, Raymond L, Stockhause M, Tedds J, Vardigan M, et al (2017) Key components of data publishing: using current best practices to develop a reference model for data publishing. *Int J Digit Libr* 18(2):77–92
- Barbosa L, Pham K, Silva C, Vieira MR, Freire J (2014) Structured open urban data: understanding the landscape. *Big data* 2(3):144–154
- Barnaghi P, Wang W, Henson C, Taylor K (2012) Semantics for the internet of things: early progress and back to the future. *Int J Semant Web and Inf Syst (IJSWIS)* 8(1):1–21
- Barnaghi P, Sheth A, Henson C (2013) From data to actionable knowledge: big data challenges in the web of things [guest editors' introduction]. *IEEE Intell Syst* 28(6):6–11
- Behkamal B, Kahani M, Paydar S, Dadkhah M, Sekhavaty E (2010) Publishing Persian linked data: challenges and lessons learned. In: 5th International Symposium on Telecommunications (IST). IEEE, Tehran. pp 732–737. <https://doi.org/10.1109/ISTEL.2010.5734119>
- Berners-Lee T, Connolly D, Swick RR (1999) Web architecture: describing and exchanging data. <https://www.w3.org/1999/04/WebData>. Accessed 10 Jan 2018
- Bhagat J, Tanoh F, Nzuobontane E, Laurent T, Orlowski J, Roos M, Wolstencroft K, Alekseyevs S, Stevens R, Pettifer S, et al (2010) Biocatologue: a universal catalogue of web services for the life sciences. *Nucleic Acids Res* 38(suppl_2):689–694. <https://doi.org/10.1093/nar/gkq394>
- Bikakis N, Sellis TK (2016) Exploration and visualization in the web of big linked data: a survey of the state of the art. In: Proceedings of the 6th International Workshop on Linked Web Data Management, Bordeaux
- Bizer C (2009) The emerging web of linked data. *IEEE Intell Syst* 24(5):87–92
- Bizer C, Heath T, Berners-Lee T (2009) Linked data—the story so far. In: Sheth A (ed). *International Journal on Semantic Web and Information Systems* Vol. 5. pp 1–22. <https://doi.org/10.4018/jswis.2009081901>
- Bizer C, Boncz P, Brodie ML, Erling O (2012) The meaningful use of big data: four perspectives—four challenges. *ACM Sigmod Rec* 40(4):56–60
- Bouquet P, Stoermer H, Vignolo M (2012) Web of data and web of entities: identity and reference in interlinked data in the semantic web. *Philos Technol* 25(1):5–26
- Brandt KS, de Boer V (2015) Linked data for the international aid transparency initiative. *J Data Semant* 4(3):187–211
- Braunschweig K, Eberius J, Thiele M, Lehner W (2012a) Open—enabling non-expert users to extract, integrate, and analyze open data. *Datenbank-Spektrum* 12(2):121–130
- Braunschweig K, Eberius J, Thiele M, Lehner W (2012b) The state of open data limits of current open data platforms. In: Proceedings of Web Science Track at the 21st International World Wide Web Conference. W3C, Lyon
- Bröring A, Echterhoff J, Jirka S, Simonis I, Everding T, Stasch C, Liang S, Lemmens R (2011) New generation sensor web enablement. *Sensors* 11(3):2652–2699
- Buranasing W, Buranarach M (2014) Publishing linked open data from semantic relation extraction for thai cultural archive. In: Proceedings of Workshop and Poster Proceedings of the 4th Joint International Semantic Technology Conference (JIST 2014). CEUR-WS, Chiang Mai. pp 85–91
- Camarda DV, Mazzini S, Antonuccio A (2012) LodLive, exploring the web of data. In: Proceedings of the 8th International Conference on Semantic Systems. ACM, New York. pp 197–200. <https://doi.org/10.1145/2362499.2362532>
- Castro LJG, McLaughlin C, Garcia A (2013) Biotea: RDFizing PubMed Central in support for the paper as an interface to the Web of Data. *J Biomed Semant* 4(1):S5
- Chang CH, Kaye M, Girgis MR, Shaalan KF (2006) A survey of web information extraction systems. *IEEE Trans Knowl Data Eng* 18(10):1411–1428
- Chen M, Mao S, Liu Y (2014) Big data: a survey. *Mob Netw Appl* 19(2):171–209
- Childs S, McLeod J, Lomas E, Cook G (2014) Opening research data: issues and opportunities. *Rec Manag J* 24(2):142–162
- Christen P (2008) Febrl: an open source data cleaning, deduplication and record linkage system with a graphical user interface. In:

- Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, New York. pp 1065–1068. <https://doi.org/10.1145/1401890.1402020>
33. Chun SA, Shulman S, Sandoval R, Hovy E (2010) Government 2.0: making connections between citizens, data and government. *Inf Polity* 15(1):1
 34. Colpaert P (2014) Route planning using linked open data. In: Proceedings of the 11th European Semantic Web Conference. Springer, Anissaras. pp 827–833. https://doi.org/10.1007/978-3-319-07443-6_56
 35. Conradie P, Choenni S (2014) On the barriers for local government releasing open data. *Gov Inf Q* 31:S10–S17
 36. Coyle K, Silvello G, Tammara AM (2014) Comparing methodologies: linked open data and digital libraries. In: Proceedings of the Third AIUCD Annual Conference on Humanities and Their Methods in the Digital Ecosystem. ACM, Bologna. p 3. <https://doi.org/10.1145/2802612.2802615>
 37. Da Silva JAT, Dobránszki J (2015) Potential dangers with open access data files in the expanding open data movement. *Publ Res Q* 31(4):298–305
 38. Aquin M (2016) On the use of linked open data in education: current and future practices. In: Mouroumsev D, d'Aquin M (eds). *Open Data for Education: Linked, Shared, and Reusable Data for Teaching and Learning*. Lecture Notes in Computer Science (9500). Springer. pp 3–15. https://doi.org/10.1007/978-3-319-30493-9_1
 39. Devarakonda R, Palanisamy G, Wilson BE, Green JM (2010) Mercury: reusable metadata management, data discovery and access system. *Earth Sci Inf* 3(1-2):87–94
 40. Dinter B, Gluchowski P, Schieder C (2015) A stakeholder lens on metadata management in business intelligence and big data—results of an empirical investigation. In: Proceedings of 21st Americas Conference on Information Systems. AIS, Fajardo. pp 439–450
 41. Dittrich KR, Jonscher D (2000) All together now: towards integrating the world's information systems. In: Proceedings of Las Jornadas de Ingeniería del Software y Bases de Datos (JISBD). Universidad de Valladolid, Valladolid
 42. Dong XL, Naumann F (2009) Data fusion: resolving data conflicts for integration. *Proc of the VLDB Endowment* 2(2):1654–1655
 43. Erkimbaev A, Zitserman VY, Kobzev G, Serebrjakov V, Teymurazov K (2013) Publishing scientific data as linked open data. *Sci Tech Inf Process* 40(4):253
 44. Esteve C, Serra L (2014) The mapping, selecting and opening of data: the records management contribution to the open data project in girona city council. *Rec Manag J* 24(2):87–98
 45. Fisher DM (2004) The business process maturity model: a practical approach for identifying opportunities for optimization. *Bus Process Trends* 9(4):11–15
 46. Freitas A, Curry E (2016) Big data curation. In: Cavanillas J, Curry E, Wahlster W (eds). *New Horizons for a Data-Driven Economy*. Springer. pp 87–118. https://doi.org/10.1007/978-3-319-21569-3_6
 47. Frey JG, Bird CL (2014) Scientific and technical data sharing: a trading perspective. *J Comput Aided Mol Des* 28(10):989–996
 48. Frischmuth P, Martin M, Tramp S, Riechert T, Auer S (2015) Ontowiki—an authoring, publication and visualization interface for the data web. *Semant Web* 6(3):215–240
 49. Gallagher J, Orcutt J, Simpson P, Wright D, Pearlman J, Raymond L (2015) Facilitating open exchange of data and information. *Earth Sci Inf* 8(4):721–739
 50. Gama K, Lóscio BF (2014) Towards ecosystems based on open data as a service. In: Proceedings of 16th International Conference on Enterprise Information Systems (ICEIS). Lisbon Vol. 2. pp 659–664. <https://doi.org/10.5220/0004974506590664>
 51. Geiger CP, Von Lucke J (2012) Open government and linked open government data. *eJournal of eDemocracy and open Gov (JeDEM)* 4(2):265–278
 52. Goble C, Stevens R, Hull D, Wolstencroft K, Lopez R (2008) Data curation+ process curation= data integration+ science. *Brief Bioinform* 9(6):506–517
 53. Goldacre B, Gray J (2016) OpenTrials: towards a collaborative open database of all available information on all clinical trials. *Trials* 17(1):164
 54. Gracy KF (2015) Archival description and linked data: a preliminary study of opportunities and implementation challenges. *Arch Sci* 15(3): 239–294
 55. Group OGW (2007) Eight principles of open government data. https://public.resource.org/8_principles.html. Accessed 10 Jan 2018
 56. Halevy A, Rajaraman A, Ordille J (2006) Data integration: the teenage years. In: Proceedings of the 32nd international conference on Very large data bases. VLDB Endowment, Seoul. pp 9–16
 57. Halper F, Krishnan K (2013) TDWI big data maturity model guide: interpreting your assessment score, The Data Warehousing Institute (TDWI), Renton
 58. Hartig O, Langegger A (2010) A database perspective on consuming linked data on the web. *Datenbank-Spektrum* 10(2):57–66
 59. Heath T, Bizer C (2011) Linked data: evolving the web into a global data space. *Synth Lect Semant Web: theory and technol* 1(1):1–136
 60. Hoxha J, Brahaj A (2011) Open government data on the web: a semantic approach. In: Proceedings of the 2011 International Conference on Emerging Intelligent Data and Web Technologies. IEEE, Tirana. pp 107–113. <https://doi.org/10.1109/EIDWT.2011.24>
 61. Hrynaszkiewicz I, Khodiyar V, Hufton AL, Sansone SA (2016) Publishing descriptions of non-public clinical datasets: proposed guidance for researchers, repositories, editors and funding organisations. *Res Integr and Peer Rev* 1(1):6
 62. Hyland B, Atemez G, Villazón-Terrazas B (2014) Best practices for publishing linked data. <https://www.w3.org/TR/2014/NOTE-ld-bp-20140109/>. Accessed 10 Jan 2018
 63. Jacobs I, Walsh N (2004) Architecture of the world wide web. <https://www.w3.org/TR/webarch/>. Accessed 10 Jan 2018
 64. Janssen M, Charalabidis Y, Zuidewijk A (2012) Benefits, adoption barriers and myths of open data and open government. *Inf Syst Manag* 29(4):258–268
 65. Kalampokis E, Tambouris E, Tarabanis K (2013) On publishing linked open government data. In: Proceedings of the 17th Panhellenic Conference on Informatics. ACM, Koblenz. pp 25–32. <https://doi.org/10.1145/2491845.2491869>
 66. Kansa EC, Kansa SW, Burton MM, Stankowski C (2010) Googling the grey: open data, web services, and semantics. *Archaeologies* 6(2): 301–326
 67. Kitchenham B, Charters S (2007) Guidelines for performing Systematic Literature Reviews in Software Engineering. Technical Report, Evidence-Based Software Engineering Project, Keele University and Durham University
 68. Kucera J (2015) Open government data publication methodology. *J Syst Integr CSSI* 6(2):52–61. <https://doi.org/10.20470/jsi.v6i2.231>
 69. Kucera J, Chlapek D, Klímek J, Necaský M (2015) Methodologies and best practices for open data publication. In: Proceedings of Annual International Workshop on Databases, Texts, Specifications and Objects (DATEO). CEUR-WS.org, Jicin. pp 52–64
 70. Labrinidis A, Jagadish HV (2012) Challenges and opportunities with big data. *Proc VLDB Endowment* 5(12):2032–2033
 71. Laender AH, Ribeiro-Neto BA, da Silva AS, Teixeira JS (2002) A brief survey of web data extraction tools. *ACM Sigmod Rec* 31(2):84–93
 72. LaValle S, Lesser E, Shockley R, Hopkins MS, Kruschwitz N (2011) Big data, analytics and the path from insights to value. *MIT Sloan Manag Rev* 52(2):21
 73. Lóscio BF, Oliveira MIS, Bittencourt II (2015) Publicação e Consumo de Dados na Web: Conceitos e Desafios. *Tópicos em Gerenciamento de Dados e Informações (Mini Cursos - SBBD 2015) d:39–69*. <http://dexl.incc.br/sbbd2015/anais/ShortCourses.pdf>. Accessed 10 Jan 2018
 74. Lóscio BF, Burle C, Calegari N (2016a) Data on the Web Best Practices. W3C Recommendation, World Wide Web Consortium (W3C). Accessed in 10 Jan 2018
 75. Lóscio BF, Burle C, Calegari N (2016b) Data on the web best practices: challenges and benefits. In: Proceedings of the 2016 Open Data Reserach Symposium. The Gov Lab, Madrid
 76. Maali F, Erickson J, Archer P (2014) Data catalog vocabulary, W3C Recommendation. <https://www.w3.org/TR/vocab-dcat/>. Accessed 12 Dec 2018
 77. Madhavan J, Jeffery SR, Cohen S, Dong X, Ko D, Yu C, Halevy A (2007) Web-scale data integration: you can only afford to pay as you go. In: Proceedings of 3rd the Conference on Innovative Data Systems Research, Asilomar. pp 342–350
 78. Martin M, Abicht K, Stadler C, Ngonga Ngomo AC, Soru T, Auer S (2015) Cubeviz: exploration and visualization of statistical linked data. In: Proceedings of the 24th International Conference on World Wide Web. ACM, Florence. pp 219–222. <https://doi.org/10.1145/2740908.2742848>

79. Matthews B, Bunakov V, Jones C, Crompton S (2013) Investigations as research objects within facilities science. In: Proceedings of the 2013 International Conference on Theory and Practice of Digital Libraries. Springer, Valletta. pp 127–140. https://doi.org/10.1007/978-3-319-08425-1_12
80. de Mendonca VRL, Rodrigues CL, de MN Soares FAA, Vincenzi AMR (2013) Static analysis techniques and tools: a systematic mapping study. In: Proceedings of the 8th International Conference on Software Engineering Advances. IARIA, Guimarães. pp 72–78
81. Milić P, Veljković N, Stoimenov L (2015) Linked relations architecture for production and consumption of linksets in open government data. In: Proceedings of Conference on e-Business, e-Services and e-Society. Springer, Delft. pp 212–222. https://doi.org/10.1007/978-3-319-25013-7_17
82. Möller K (2013) Lifecycle models of data-centric systems and domains: the abstract data lifecycle model. *Semant web* 4(1):67–88. <http://dl.acm.org/citation.cfm?id=2595053.2595060>
83. Moßgraber J, Hilbring D (2014) Automating the web publishing process of environmental data by using semantic annotations. In: Proceedings of International Conference on Multimedia Retrieval (ICMR). CEUR-WS.org, Glasgow. pp 1–6
84. Mouhoub ML, Grigori D, Manouvrier M (2014a) A framework for searching semantic data and services with SPARQL. In: Proceedings of the 12th International Conference on Service-Oriented Computing. Springer, Berlin. pp 123–138. https://doi.org/10.1007/978-3-662-45391-9_9
85. Mouhoub ML (2014b) Searching linked data and services with a single query. In: Proceedings of the 11th European Semantic Web Conference. Springer, Anissaras. pp 855–863. https://doi.org/10.1007/978-3-319-07443-6_59
86. Mouzakitis S, Papaspyros D, Petychakis M, Koussouris S, Zafeiropoulos A, Fotopoulou E, Farid L, Orlandi F, Attard J, Psarras J (2017) Challenges and opportunities in renovating public sector information by enabling linked data and analytics. *Inf Syst Front* 19(2):321–336
87. Murphy F, Dallmeier-Tiessen S, Bloom T, Nurnberger A, Austin CC, Tedds J, Whyte A, Raymond L, Stockhouse M, Vardigan M, Khodiyar V (2015) WDS-RDA-f11 publishing data workflows WG synthesis final corrected. <https://dor.org/10.5281/zenodo.33899>. <http://zenodo.org/record/33899>
88. Nath K, Iswary R (2015) What comes after web 3.0? Web 4.0 and the future. In: Proceedings of the International Conference and Communication System (ICCS'15). Shillong, India. pp 337–341
89. Oliveira MIS, Lóscio BF (2018) What is a data ecosystem? In: Proceedings of the 19th International Digital Government Research Conference on Digital Government Research. ACM, Delft. <https://doi.org/10.1145/3209281.3209335>
90. Oliveira MIS, Oliveira LA, Batista MGR, Lóscio BF (2018a) Towards a meta-model for data ecosystems. In: Proceedings of the 19th International Digital Government Research Conference on Digital Government Research. ACM, Delft. <https://doi.org/10.1145/3209281.3209333>
91. Oliveira OMIS, Lairson Alencar, Santos WCdR, Lóscio BF (2018b) Data on the web management system: a reference model. In: Proceedings of the 19th International Digital Government Research Conference on Digital Government Research. ACM, Delft. <https://doi.org/10.1145/3209281.3209355>
92. Peña KIC (2015) Comparative analysis of public policies in open access models in latin america. brazil and argentina cases. *Int J Educ Technol High Educ* 12(1):15–24
93. Petticrew M, Roberts H (2008) *Systematic reviews in the social sciences: a practical guide* (1st Edition). Wiley, Hoboken
94. Petychakis M, Vasileiou O, Georgis C, Mouzakitis S, Psarras J (2014) A state-of-the-art analysis of the current public data landscape from a functional, semantic and technical perspective. *J Theor Appl Electron Commerce Res* 9(2):34–47
95. Piedra N, Tovar E, Colomo-Palacios R, Lopez-Vargas J, Alexandra Chicaiza J (2014) Consuming and producing linked open data: the case of Opencourseware. *Program* 48(1):16–40
96. Pollock R (2011) Building the (open) data ecosystem. <https://blog.okfn.org/2011/03/31/building-the-open-data-ecosystem/>. Accessed 2 July 2017
97. Prud E, Seaborne A, et al (2008) SPARQL query language for RDF. <https://www.w3.org/TR/rdf-sparql-query/>. Accessed 21 Apr 2017
98. Rademaker A, Oliveira DAB, de Paiva V, Higuchi S, e Sá AM, Alvim M (2015) A linked open data architecture for the historical archives of the Getulio Vargas Foundation. *Int J Digit Libr* 15(2-4):153–167
99. Radulovic F, Poveda-Villalón M, Vila-Suero D, Rodríguez-Doncel V, García-Castro R, Gómez-Pérez A (2015) Guidelines for linked data generation and publication: an example in building energy consumption. *Autom Constr* 57:178–187
100. Rodríguez-Iglesias A, Rodríguez-González A, Irvine AG, Sesma A, Urban M, Hammond-Kosack KE, Wilkinson MD (2016) Publishing fair data: an exemplar methodology utilizing PHI-base. *Front Plant Sci* 7:641. <https://doi.org/10.3389/fpls.2016.00641>
101. Dulong de Rosnay M, Janssen K (2014) Legal and institutional challenges for opening data across public sectors: towards common policy solutions. *J Theor Appl Electron Commerce Res* 9(3):1–14
102. Russom P (2013) *Managing big data*. TDWI Best Practices Report, The Data Warehousing Institute (TDWI), Renton
103. dos Santos Brito K, da Silva Costa MA, Garcia VC, de Lemos Meira SR (2014) Brazilian government open data: implementation, challenges, and potential opportunities. In: Proceedings of the 15th Annual International Conference on Digital Government Research. ACM, Aguascalientes. pp 11–16. <https://doi.org/10.1145/2612733.2612770>
104. Shadbolt N, O'Hara K (2013) Linked data in government. *IEEE Internet Comput* 17(4):72–77
105. Shah RC, Kesan JP, Kennis A (2008) Lessons for government adoption of open standards: a case study of the massachusetts policy. *J Inf Technol Polit* 5(4):387–398
106. Shaw M (2003) Writing good software engineering research papers. In: Proceedings of the 25th International Conference on Software Engineering. IEEE, Portland. pp 726–736
107. Sheth A, Henson C, Sahoo SS (2008) Semantic sensor web. *IEEE Internet Comput* 12(4)
108. Stonebraker M, Bruckner D, Ilyas IF, Beskales G, Cherniack M, Zdonik SB, Pagan A, Xu S (2013) Data curation at scale: The data tamer system. *CIDR*
109. Villazo?n-Terrazas B, Vila-Suero D, Garijo D, Vilches-Blazquez L, Poveda-Villalón M, Mora J, Corcho O, Gómez-Pérez A (2012) Publishing linked data - there is no one-size-fits-all formula. Proceedings of the European Data Forum 2012. <http://oa.upm.es/14465/>, oeg
110. van der Waal S, Wecler K, Ermilov I, Janev V, Milošević U, Wainwright M (2014) Lifting open data portals to the data web. In: *Linked Open Data—Creating Knowledge Out of Interlinked Data*. Springer. pp 175–195
111. Wang C, Wang Q, Ren K, Cao N, Lou W (2012) Toward secure and dependable storage services in cloud computing. *IEEE Trans Serv Comput* 5(2):220–232
112. Willighagen EL, Waagmeester A, Spjuth O, Ansell P, Williams AJ, Tkachenko V, Hastings J, Chen B, Wild DJ (2013) The ChEMBL database as linked open data. *J Cheminformatics* 5(1):23
113. Yamamoto Y, Yamaguchi A, Yonezawa A (2013) Building Linked Open Data towards integration of biomedical scientific literature with DBpedia. *J Biomed Semant* 4(1):8
114. Yang TM, Wu YJ (2016) Examining the socio-technical determinants influencing government agencies' open data publication: a study in taiwan. *Gov Inf Q* 33(3):378–392
115. Zhang J, Dawes SS, Sarkis J (2005) Exploring stakeholders' expectations of the benefits and barriers of e-government knowledge sharing. *J Enterp Inf Manag* 18(5):548–567
116. Zhao J (2010) Publishing chinese medicine knowledge as linked data on the web. *Chin Med* 5(1):27
117. Zikopoulos P, Eaton C (2011) *Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data* (1st Edition). McGraw-Hill Osborne Media, New York City
118. Zuiderwijk A, Janssen M, Choenni S (2012a) Open data policies: impediments and challenges. In: 12th European conference on e-government (ECEG 2012), Barcelona, Spain. pp 794–802
119. Zuiderwijk A, Janssen M, Choenni S, Meijer R, Alibaks RS (2012b) Socio-technical impediments of open data. *Electron J e-Gov* 10(2):156–172