International Journal on Semantic Web and Information Systems
Volume 13 • Issue 4 • October-December 2017

The Use of Software Tools in Linked Data Publication and Consumption:
A Systematic Literature Review

Armando Barbosa, Federal University of Alagoas, Maceió, Brazil
Ig Ibert Bittencourt, Federal University of Alagoas, Maceió, Brazil
Sean Wolfgand Matsui Siqueira, Federal University of the State of Rio de Janeiro (UNIRIO), Rio de Janeiro, Brazil
Rafael de Amorim Silva, Federal University of Alagoas, Maceió, Brazil
Ivo Calado, Federal Institute of Alagoas, Rio Largo, Brazil

ABSTRACT

To reduce the complexity intrinsic to LD manipulation, software tools are used to publish or consume data associated to LD activities. However, few developers have a broad understanding of how software tools may be used in publication or consumption of Linked Data. The goal of this work is to investigate the use of software tools in Linked Data publication and consumption processes. More specifically, understanding how these software tools are related to process of publication or consumption of LD. In order to meet their goal, the authors conducted a Systematic Literature Review (SLR) to identify the studies on the use of software tools in these processes. The SLR gathered 6473 studies, of which only 80 studies remained for final analysis (1.25% of the original sample). The highlights of the study are: (1) initial steps of the publication process are fairly supported by the software tools; (2) Non-RDF serialization is fairly supported in publication and consumptions process by the software tools; and (3) there are non-supported steps in consumption and publication processes by the tools.

KEYWORDS

Consumption Process, Linked Data, Ontology, Publication Process, Software Tools

1. INTRODUCTION

In last decade, the evolution of the Web has increased the consumption and production of content. This growth implies in several challenges such as a longer time required to perform searches or browsing, data consistency problems, and misinterpretations. The users and machines perceive these challenges, as they are necessary to access a set of web pages to find relevant information. In addition, most of the data available on the Web are unstructured (images, blogs, web content, etc.), which means that machines are not able to automatically understand such data content. Therefore, it is essential to provide alternatives for structuring data in a machine understandable manner.

Linked Data (LD) is an effective option to structure such data, presenting a set of best practices for guiding publishing and connecting these structured data on the Web (Hyland & Wood, 2011). LD adopts a generic abstract data model for describing resources called Resource Description Framework

DOI: 10.4018/IJSWIS.2017100104

(RDF). This framework is responsible for describing terms and their relationships, by using triples of subjects, predicates and objects. Additionally, ontologies may provide enrichment to Linked Data by adding semantics to their RDF statements (Heath & Bizer, 2011). Moreover, LD brings some benefits which may be applied in multiple research fields, such as data interoperability and terms definition (Bauer & Kaltenböck, 2011). Some applications of LD include Open Government (Ding, Lebo, Erickson, DiFranzo, Williams, Li, & Hendler, 2011; Shadbolt, O'Hara, Berners-Lee, Gibbins, Glaser, & Hall, 2012) (e.g. to improve of transparency and provenance), Education (Dietze, Yu, Giordano, Kaldoudi, Dovrolis, & Taibi, 2012) (e.g. to improve education repositories and content sharing), and Business (Barnes & Martin, 2002) (e.g. to improve business opportunities and business value).

LD is defined by both publication and consumption processes. Recently, W3C (2014) has published a list of 10 steps to fulfill LD publication requirements: (1) to prepare stakeholders; (2) to select dataset for reuse; (3) to model the data to represent data objects and their relationships; (4) to specify a suitable open data license; (5) to define URI for linked data; (6) to describe objects with standard vocabularies; (7) to convert data to a linked data representation; (8) to provide machine access to data; (9) to announce to the public; and (10) to establish a social contract of a linked data publisher. On the other hand, Bauer and Kaltenböck (2011) have published a list of 7 steps related to LD consumption process: (1) to specify concrete use cases for new services or applications; (2) to evaluate relevant data sources and datasets; (3) to check licenses for use and reuse provided by data owners; (4) to create consumption patterns to specify what data are reused from a certain data source; (5) to manage alignment, caching and update mechanisms; (6) to create mashups and applications to provide user-friendly graphical user interfaces (GUIs) and powerful services for end users; and (7) to establish sustainable partnerships.

Both aforementioned processes are essentials to improve data quality, interoperability, and discoverability of data, as well as to support the development of rich applications. Additionally, considering the number of steps and activities, it is important the development of tools for supporting Linked Data publication and consumption processes. Considering the huge number of Linked Data-related tools proposed recently (Bizer, Heath, & Berners-Lee, 2009), it is a time-consuming task, for publishers or consumers, to know the available features of each tool in order to make properly decision about software tools that should be used according to each step. For this reason, a high-level question was raised: Which software tools have been used to support the publication and/or consumption of Linked Data? In this context, our purpose is to identify what tools have been used to perform the aforementioned tasks, providing a broad vision about Linked Data software tools to publish and/or consume linked data. Additionally, we use the categories provided by W3C[1] to classify the software tools according to their features.

There are some similar works that investigated the use of software tools to publish or consume Linked Data (Corlosquet, Delbru, Clark, Polleres, & Decker, 2009; Bizer et al., 2009). However, these studies do not capture all the aspects and evidences that we are interested, such as support and formats. Moreover, their work is somewhat limited, as they do not systematically review the literature to investigate the works that jointly cover software tools and Linked Data Publication and Consumption Process.

Therefore, in this paper we use the Systematic Literature Review (SLR) method to understand which software tools support publication and consumption of linked data. We utilize the method defined in (Kitchenham & Charters, 2007) to identify, evaluate, interpret and synthesize the available studies in order to answer particular research questions on the symbiosis of software tools for publication and/or consumption of linked data, and to establish the state of evidence with an in-depth

analysis. Thought the analysis we generate six questions. These questions may guide new research line, software solutions, and discussion.

The rest of this paper is organized as follows: Section 2 describes the SLR method used in this review and primary results. Section 3 presents the quality assessment. Section 4 presents results. Section 5 presents a discussion about the results. Section 6 presents threats to validity. Finally, in the last section we present the conclusions and future work of this SLR.

2. SYSTEMATIC LITERATURE REVIEW

A systematic literature review (SLR) demonstrates the state of the art in some particular research field, evaluating findings related to a given research issue, a topic area, or even a phenomenon. It is worth to differentiate a SLR from a Systematic Mapping. According to Kitchenham et al. (2007), Systematic Mappings or Scoping Studies provide a wide overview of a research area. The results of a mapping study can identify areas suitable to conducting a SLR and areas where a primary study is more appropriate.

In this work, we adopt the guidelines proposed by Kitchenham et al. (2007) to conduct our SLR. We also adopt the tool called StArt (State of the Art through Systematic Reviews) (LaPES, 2012) to conduct our SLR, reducing errors inherent to the SLR process. This tool is validated in (Hernandes, Zamboni, Fabbri, & Thommazo, 2012), obtaining a high performance on the execution of a SLR.

In this work, we investigate the following research Main Question (MQ):

MQ → Which software tools support the publication and/or consumption of Linked Data?

According to Kitchenham et al. (2007), the main step for any systematic review is the definition of the research questions, since they are responsible for guiding the whole review process. Based on such assumption, and similarly to other systematic reviews (Dermeval, Vilela, Bittencourt, Isotani, & Brito, 2014), we elaborate three research questions in order to enable us to answer the MQ.

- **RQ1** → *Which are the linked data-based publication / consumption steps utilized by software tools in current literature?*
- **RQ2** → *Which serialization formats does the tool support?*
- **RQ3** → *Is the usage of software tool empirically evaluated?*

RQ1 provides a starting point to understand the software tools and the main approach (publish/consume) supported by them. This question is responsible for identifying the steps of the publication and of the consumption process supported by the software tools. RQ2 identifies the main serializable formats (RDF, OWL, JSON, CSV, XML and others) used for publishing or consuming linked data. Finally, RQ3 intends to analyze if such studies provide an empirical evaluation of the software tools. Besides, Table 1 presents the central motivation for each research question.

The step II defines inclusion and exclusion criteria adopted in this work. According to Kitchenham et al. (2007), criteria in the SLR process identify those papers that provide direct evidence about research questions, reducing a possible bias. We adopted two types of criteria in the SLR: (1) inclusion criteria; and (2) exclusion criteria. The former defines all the necessary requirements in order an article must be considered in the SLR process. The latter defines all the requirements for an article not to be included in this process. We considered primary papers (those that present some proposal in the area) in the inclusion criteria whereas secondary papers (i.e. those that only review an area topic such as a SLR) in the exclusion criteria. Regarding to publication date, studies are eligible for inclusion in the SLR process whether they were peer-reviewed and published from January 2006

Table 1. Research Question Motivation

RQ	Motivation	Research Question
RQ1	In general, tools support only some steps of the publication/ consumption process. In this context, it is important to determine for each analyzed tool the exact set of supported steps.	Which are the linked data-based publication / consumption steps utilized by software tools in current literature?
RQ2	Considering there are different serialization formats available, we want to determine the formats that each analyzed tool supports.	Which serialization formats do the tool support?
RQ3	The evaluation of a publication / consumption tool is essential to quantify the effectiveness of the analyzed tool. So, we intend to verify which empirical evaluations were applied for each tool.	Is the usage of software tool empirically evaluated?

until December 2015. In the other hand, a paper is excluded from the process whether it belongs to some exclusion criteria defined in Table 2.

To perform the step III, we considered the following electronic databases:

- Science Direct (http://www.sciencedirect.com);
- ISI Web of Knowledge (http://apps.webofknowledge.com);
- Scopus (htttp://www.scopus.com);
- Springer Link (http://link.springer.com/);
- ACM Digital Library (http://dl.acm.org);
- IEEE Xplore (http://ieeexplore.ieee.org);
- Engine Village (http://www.engineeringvillage.com).

Figure 1 shows the systematic review process and the number of papers identified at each stage.

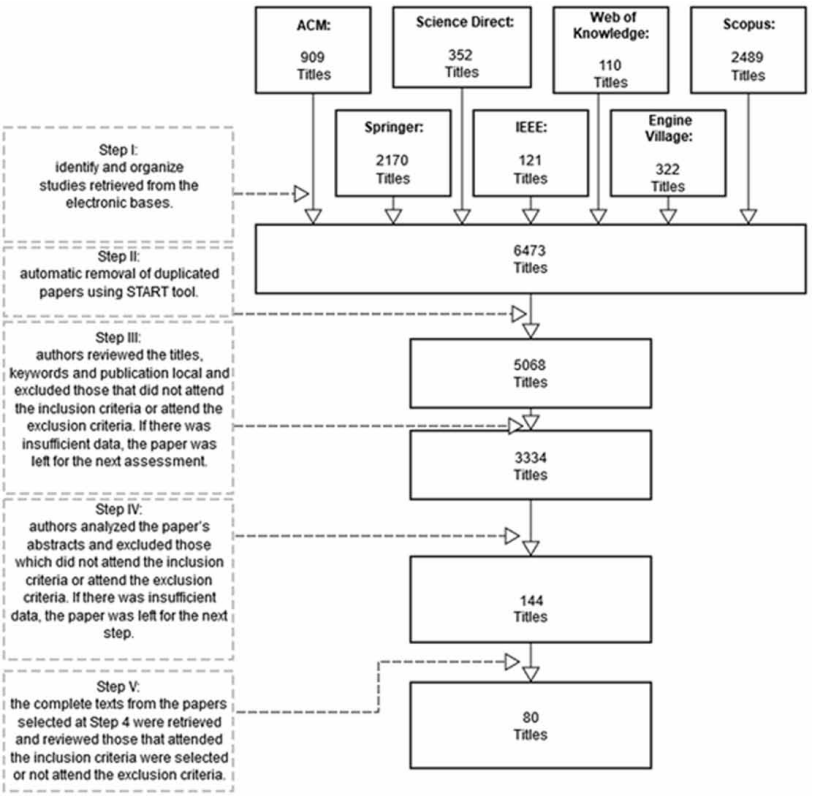
It is worth highlighting the steps related to the paper selection process. Thus, at the Step I, We extracted the papers from electronic databases using the following search terms: (1) software tool terms, containing "Environment" OR "Authoring" OR "Architecture" OR "Framework" OR "Software" OR "Application" OR "Program"; (2) linked data terms, containing "Linked Data" OR "Linked Open Data"; and (3) ontology terms, containing "Ontology" OR "Ontologies". From the search results (6,473 papers), we automatically accessed, downloaded and organized with aiding of the StArt tool. These terms were combined as follows:

(1) AND (2) AND (3)

Table 2. Inclusion and exclusion criteria

#	Inclusion Criteria	#	Exclusion Criteria
1	Papers about tools for publishing and/or consuming linked data	1	Papers in other language than English
2	Primary Studies	2	Short Papers
3	Computer Science Papers	3	Duplicated Studies
4	Only Papers in English	4	Secondary Studies
5	Papers between Jan. 2006 and Dec. 2015	5	Papers only about tools
6	Papers which answer the research questions	6	Papers only about linked data
		7	Papers only about software tools focused open data (without Linked Data)

Figure 1. Paper selection flow chart



At the Step II, we detected and removed duplicated papers using the StArt tool, remaining a set of 5,068 papers.

At the Step III, we reviewed titles and keywords of each paper and excluded those not related to the research questions (1,733 papers). The remaining set of papers (3,335) were considered for the next step since data was insufficient to decide on removal.

At the Step IV, we analyzed the abstracts and excluded papers according to the exclusion criteria (3,190 papers). Similarly to Step III, the papers with insufficient information to be classified on the exclusion criteria remained in the list to be analyzed in the next Step.

At the Step V, we retrieved the complete text from the remained papers (144), and read both introduction and conclusion sections and, additionally, performed a full-text screened of each one of the selected papers. From this step, 64 papers were excluded in this step according to the exclusion criteria.

After finishing Step V, 80 papers remained to the data extraction process, which has two activities: (1) reading the full-text of each paper and (2) saving extracted data (i.e. in a spreadsheet, database, text file, etc.). Therefore, all remaining papers of Step V were submitted for full-text reading. To save the data, we used StArt tool. Additionally, during this stage, we extracted data according to an extraction form (see Table 6 in Appendix A), which enabled us to record details about the reviewed papers.

3. QUALITY ASSESSMENT

The quality assessment of the selected studies is useful to increase the accuracy of the data extraction results. In addition, it is important to determine the validity of the inferences and in ascertaining

credibility and coherent synthesis of results. Table 3 presents the quality assessment (Kitchenham et al., 2007) criteria. Criteria 2, 4 and 8 only receive Yes (Y) or No (N) as score. Additionally, we allowed only Yes (Y) and Partially (P) scores to the last criteria since we are interested in the possible benefits to industry.

The raw data of the quality assessment is presented in Table 7[2] of the Appendix B, according to assessment question described in Table 3. The scores of 11 papers are less than 50% and the average score is 7.08 see Figure 2. Regarding the score of proposed criteria, Q1, Q2 and Q3 received the highest averages scores (>0.9), while Q4, Q5, Q8, Q9, Q10 and Q11 received intermediate scores between 0.55 and 0.83. Criteria Q6 and Q7 received the lowest average score, 0.14 and 0.13 respectively.

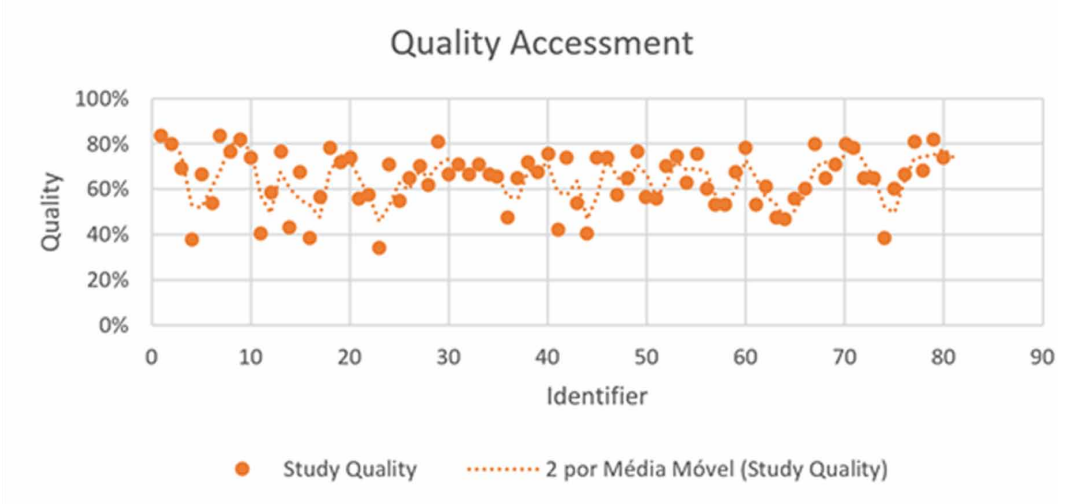
4. RESULTS

This section presents the results obtained by running the SRL method previously described. We provide a detailed overview of general features of these studies: source types, year of publication, and application context. The reviewed papers were published between January of 2006 and December of 2015. Additionally, the distribution of selected studies over publication sources, including publication name, type, count (i.e. the number of selected studies from each source), and the percentage of selected studies are available[3]. The 80 selected studies are distributed in 49 publication sources, suggesting that linked data tools have been a widespread concern in the community. The studies included in this review belong the following categories: (1) journal; (2) conference; (3) workshop; or (4) book chapters. We identified that most studies belong to the category (2) (38.75% or 31 studies), followed by category (4) (27.5% or 22 studies), category (3) (17.5% or 14 studies) and category (1) (16.25% or 13 studies). We also observed that most of these selected works were published in 2012 (22.5%)

Table 3. Quality assessment criteria

#	Questions	Possible Answers
Q1	Is there a rationale for why the study was undertaken?	Y = 1, N = 0, P = 0.5
Q2	Is the paper based on research?	Y = 1, N = 0
Q3	Is there a clear statement of the research goals?	Y = 1, N = 0, P = 0.5
Q4	Does the study reuse some framework (software, source code)?	Y = 1, N = 0
Q5	Is there an adequate description of the context in which the research was carried out?	Y = 1, N = 0, P = 0.5
Q6	How many steps of the publication process does the study support?	$\frac{\text{Number of publication steps}}{\text{Total of publication steps}}$
Q7	How many steps of the consumption process does the study support?	$\frac{\text{Number of consumption steps}}{\text{Total of consumption steps}}$
Q8	Was the study empirically evaluated?	Y = 1, N = 0
Q9	Is there a discussion about the results of the study?	Y = 1, N = 0, P = 0.5
Q10	Are the limitations of this study explicitly discussed?	Y = 1, N = 0, P = 0.5
Q11	Does the research also add value to the industrial community?	Y = 1, P = 0.5

Figure 2. Overview of the studies. Paper quality assessment distribution.



followed by 2010 (15%), 2013 (15%), 2014 (13.8%), 2011 (12.7%), 2009 (6.3%), 2008 (3.8%), 2000 (1.3%), 2007(1.3%).

The question RQ1 identifies the purpose of each software tool. We classified the tools into two categories: publication and consumption (see Figure 4). The software tools can be classified in one or both categories, according to the supported steps. Figure 3 depicts that 9 software tools (11.11%) provide support to one or more steps of both linked data consumption and publication processes. These tools can be classified, under Publication point of view, according to the following features: RDF publisher (S01, S19, S25, S37, and S38), RDF synchronizer (S33), RDF compactor (S18), Linked Data consulting language (S21) and search federated engine (S27). Table 4 describes details such as source-code language, platform, publisher, supported serializable formats, evaluation method and publication year. Figure 4 depicts the distribution of these 9 studies, in publication and consumption processes. Most solutions (77.77%; 7 studies) are focused on providing machine access (endpoint, publishing data sets, etc.) to data (SPARQL, RDF and RDFa). On the other hand, only one software tool (11.11%, 1 study) is concerned with the conversion of structured data to linked data and another one (11.11%) is concerned with three steps: definition of URIs for Linked Data, announcement of data sets and to establish social contract of a linked data publisher.

The question RQ2 identifies serializable formats supported by software tools. Figure 5 shows the relation between serialization formats and publication, as well as consumption process. We can note the growing support to RDF serialization formats, where the greatest amount of software tools related to publication process focuses on providing machine access (27 studies). On the other hand, regarding to the consumption process, the consumption patterns (28 studies) and alignment (28 studies) are the two main steps handled. It is also important to pay attention to the low support to non-RDF serializations in Step vii (Convert data) of the LD publication process. In Figure 5, it is worth to note the support is concentrated in the steps 4 and 5 of the consumption process. In fact, the most supported serializations are RDF (33.75%, 27 studies) and SPARQL (12.5%, 10 studies). On the other hand, the most supported non-RDF are XML (8.75%, 7 studies), CSV (7.5%, 6 studies) and JSON (5%, 4 studies).

The question RQ3 enables the identification of evaluated-empirically software tools. The classification of these evaluations is obtained from each paper. The number of studies for each category is shown in Figure 5. According to the results, 40% of the studies (32 studies) were not evaluated, and the other studies are distributed between experiment (36.25%, 29 studies), use case (11.25%,9

Figure 3. Classification of software tools in publication and consumption

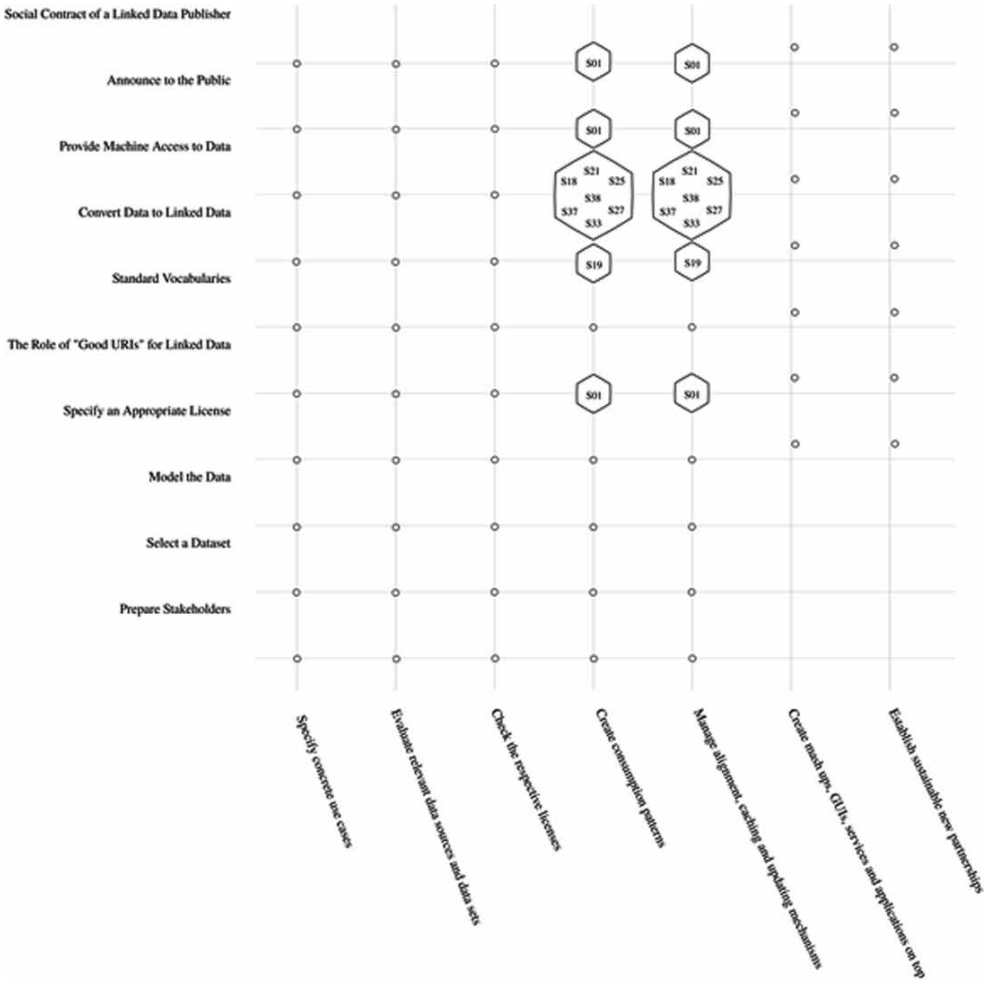


Figure 4. Software tools distribution in steps of publication and consumption processes

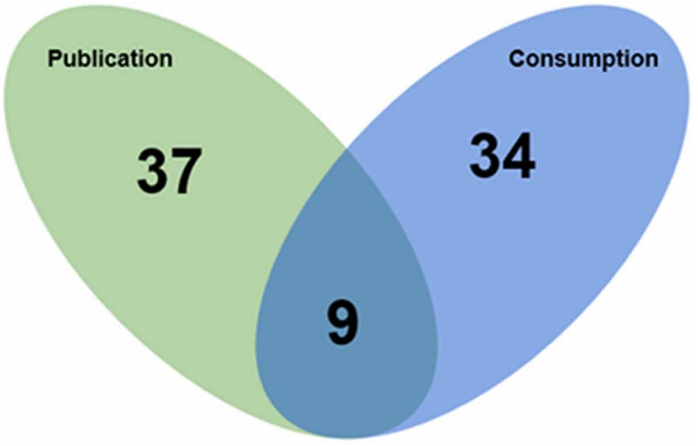


Table 4. Publication and consumption software tools

Study	Name	Developer	Year	Platform	Language	Supported Serializable Formats	Evaluation Method
S01	OrganiK project	Industry	2010	WEB	PHP	RDF	Experiment
S18	RDSZ	Academy	2014	Not described	Python	RDF	Experiment
S19	RDF extension for DRUPAL	Industry	2000	WEB	PHP	RDFa, RDFS, OWL	Does not present empirical evaluation
S21	FREyA	Industry	2012	Not described	FREyA	SPARQL, RDF	Experiment
S25	LinkedLab	Academy	2011	WEB	Not described	RDF, OWL, XML, JSON, CSV	Use Case
S27	KonneXSALT	Academy	2008	WEB	Java	RDF, SPARQL	Does not present empirical evaluation
S33	RDFSync	Industry	2007	Not described	Not described	RDF, SPARQL	Experiment
S37	Rhizomer	Academy	2011	WEB	Java, HTML, JavaScript	SPARQL	Does not present empirical evaluation
S38	Bioqueries	Academy	2011	WEB	PHP	RDF, SPARQL, RDFS	Does not present empirical evaluation

studies), proof of concept (2.5%,2 studies), benchmark (1.25%,1 study) and simulation (1.25%,1 study). Figure 7 depicts how the evaluations are distributed along publication and consumption. It is worth to note that 50% (23 studies) of the works dealing with the publication process were not evaluated.

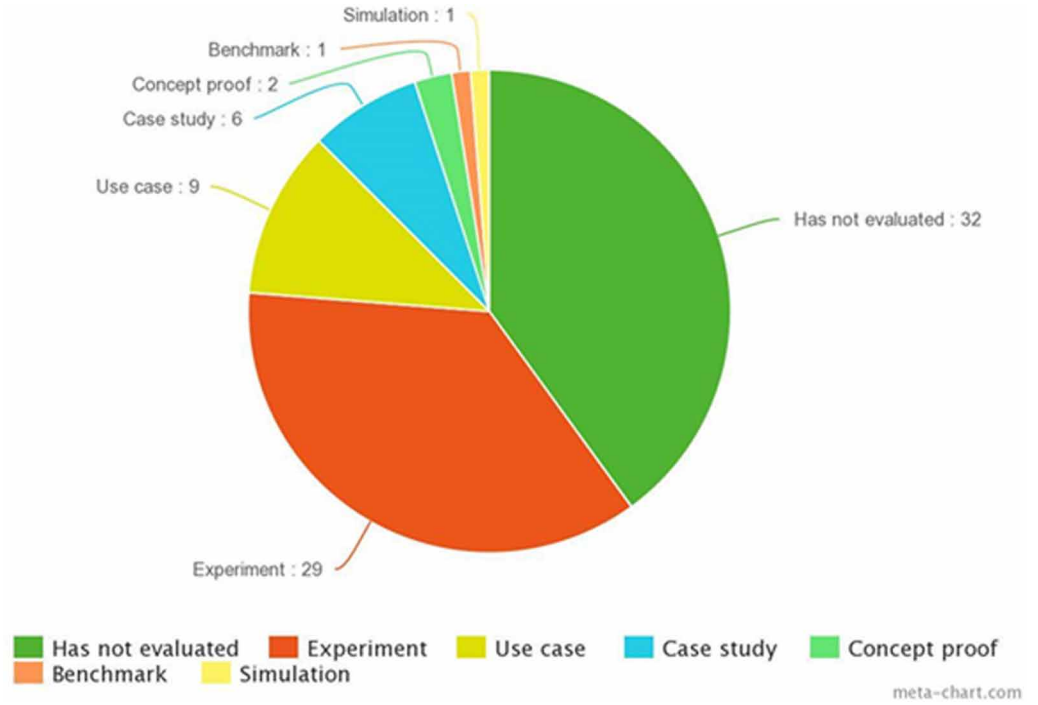
5. DISCUSSION

This section presents a discussion on the obtained results. We intend to provide our perspective about each research question, highlighting the faults in the current state of software tools for publishing and consuming Linked Data.

The results indicate an uneven distribution of software tools among the publication steps. We note that 63.82% of the studies were focused on providing machine access, followed by data conversion to linked data and the definition of URIs for linked data (with 34.04% each). On the other hand, we have detected a gap on solutions to prepare stakeholders, support open data license specification (2.12%, one study), select a dataset for reuse (10.63%, five studies), use standards vocabularies for describing objects (12.76%, six studies) and model the data to represent data objects and their relationships (24.89%, seven studies). As we can see, most of the proposed software tools support only the final steps of the publication process (44 studies) while only 24 studies support the initial steps.

The results are partially explained by the high-level vision of the publication process details provided by W3C documents. Although W3C (2014) points out the importance to use/determine data

Figure 5. Evaluation distribution



license, it lacks discussing details on how to implement such task or which vocabulary to adopt. As none analyzed tool covers the initial steps of the publication process, there is no tool able to fulfill the whole publication process. Based on the aforementioned discussion, we can point out two questions that can be raised for a research agenda.

1. How to develop software tools to cope with the initial phases of the publication process of Linked Data?
2. How to provide an integrated solution to simplify the publication of Linked Data?

Regarding the LD consumption process, the present results indicate that consumption software tools are concentrated in creating consumption patterns to specify what data are reused from a certain data source (85%, 34 studies), followed by managing alignments (85%, 34 studies). Further, the specification of use cases for new services or applications (2.5%, 1 study) and data sources and datasets evaluation (5%, 2 studies) are supported by few studies, while the remaining steps were not covered by the reviewed studies. Since developers and companies usually have specific objectives to reuse a dataset, they develop their own consumption patterns and alignments. Consequently, we can draw a third question for a research agenda:

3. How to develop and share reusable consumption patterns to facilitate the consumption of Linked Data?

In addition, it is important to provide good URIs[4], convert data to linked data, and provide machine access in publication process. Considering that these steps support a higher quantity of serialization formats, the results point out a disperse use of serializable formats in both processes. Thus, the results indicate that the support of serializable formats impacts directly on how institutions publish their data, as well as consumers reuse the published data on the Web. Another important

point is the low support provided by the software tools to previous and subsequent steps of "to create consumption patterns to specify what data are reused from a certain data source" and "to manage alignment, caching and update mechanisms".

Similarly, to the publication process, we were not able to determine any tool capable of supporting the whole consumption process. Thus, it brings us an important challenge that we summarize in the following question to the research agenda:

4. How to provide an integrated solution to simplify the consumption of Linked Data?

The creation of consumption patterns and the management of alignment mechanisms are the steps of LD consumption process that more support the serialization formats. It is worth to note the two rectangles presented in Figure 6, since they contain the highest number of supported steps and serializable formats. Therefore, the low support for serialization formats impacts negatively in the

Figure 6. Relation between publication and consumption processes and serializable formats

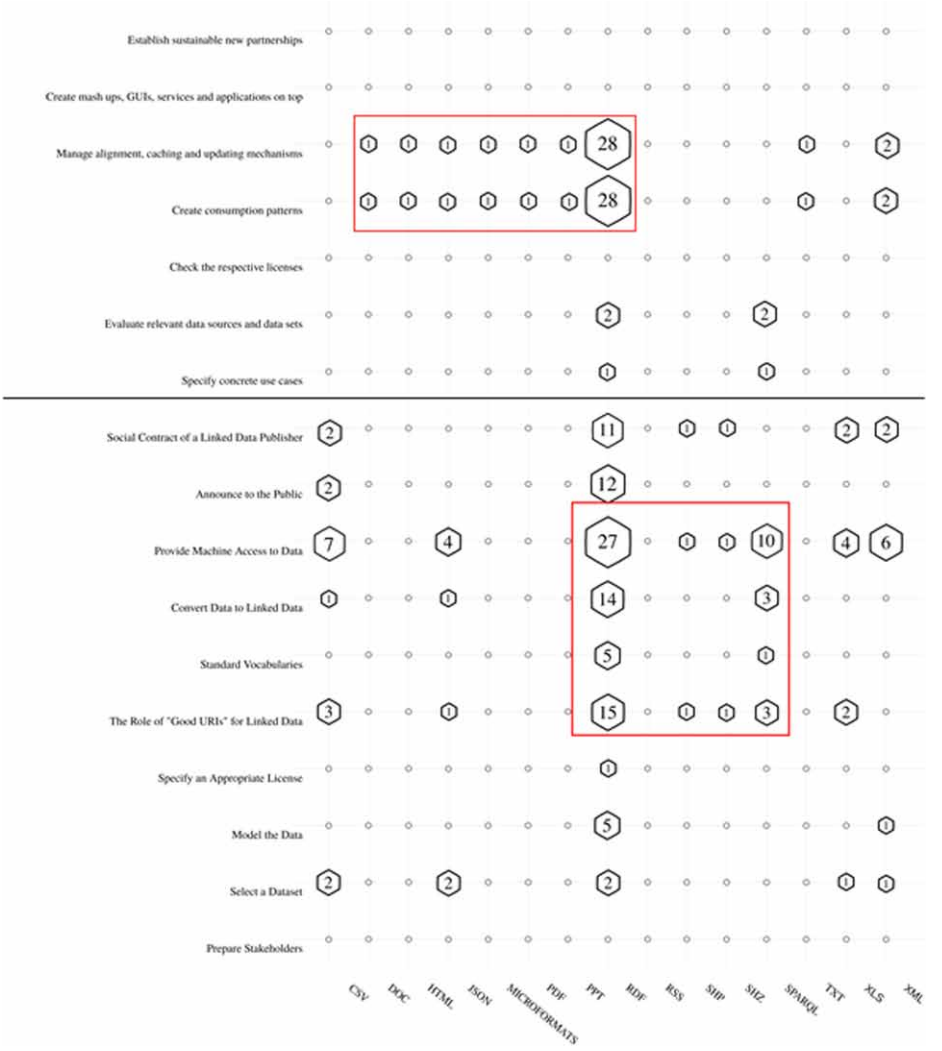
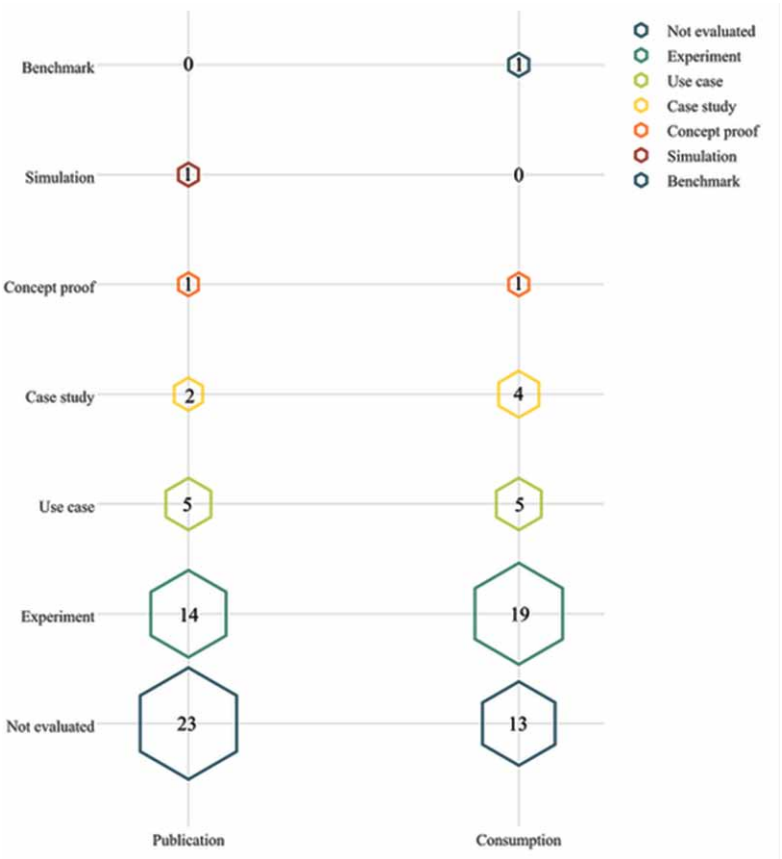


Figure 7. Evaluation by purpose



consumption process, thus it is necessary to propose approaches to increase the number of supported formats (structured and RDF serialization). This leads us to a fifth question:

5. How to provide a solution to supports RDF and non-RDF serialization formats in Linked Data Publication and Consumption process?

The empirical evaluation of any tool is essential to ensure quality, as well as to highlight the tool features. Nevertheless, almost half of evaluated tools (45%, 36 studies) do not present any kind of evaluation (empirical or not). Therefore, it is important to quantify the effectiveness of each analyzed tool to perform the intended task properly, which brings us to the last question:

6. How to validate the effectiveness of the proposed studies according to supported steps?

In the big picture, we can point out an increasing, albeit badly distributed, offer of software tools to supply the publication and consumption of Linked Data. In addition, a considerable part of these tools does not consider existing production and consumption processes during development. Thus, it is important to provide a new set of software tools that take into account the current Linked Data state of the art.

Throughout this study, we presented a considerable number of tools to assist the publisher in both processes of publishing and consuming linked data. Though the important contributions presented, according to Bizer et al. (2009) the following questions are still open: 1) the support of licenses (Heath & Bizer, 2011; Schmachtenberg, Bizer & Paulheim, 2014), 2) the automatic evaluation of data quality (Maα & Qi, 2013; Moss, Corsar, & Piper, 2012), and 3) the data alignment (Bleiholder & Naumann, 2009; Homoceanu, Kalo, & Balke, 2014). Besides, it is noteworthy the need of an integration environment to simplify the process execution.

6. THREATS TO VALIDITY

As threats to the validity of this SLR, the main constructs in this review are the three concepts "Software Tools", "Linked Data" and "Ontologies". For the first concept, we used the terms "environment", "authoring", "architecture", "framework", "software", "application" and "program" to collect studies related to software tool approaches. For the second concept, the terms "Linked Data" and "Linked Open Data" were used to cover the potentially relevant studies on Linked Data from the database search. For the third concept, the terms "ontology" and its plural "ontologies" were used to collect studies related to ontology approaches. However, we did not use control papers during the search on digital libraries, which may be considered as the main threat to validity. Additionally, we did not perform a complementary manual search.

7. CONCLUSION

In this work, we conducted a SLR to investigate the support of software tools to publish and consume Linked Data. Our goal was to improve the understanding of how these software tools support each step of both processes.

Eighty studies were fully investigated, of which nine provide support in one or more steps of the consumption and publication processes. In regards to the publication process, it is possible to evidence the low support in the initial steps. Moreover, in the consumption process, it is possible to evidence the software tools focus the support of the development of consumption patterns, as well as data alignment.

As previously discussed, we raised a set of questions, which involves the steps in processes, software tools, serialization formats and the relations among them. Finally, we highlighted some important questions that are revised and summarized in Table 5 in order to suggest a research agenda on linked data.

Based on (Kitchenham et al., 2007) we can highlight some contributions provided by this study. Such contributions include:

- The identification of gaps in current research in order to suggest areas for further investigation;
- The provision of a framework/background in order to properly position new research activities;
- The summarization of existing evidence, benefits and limitations.

The result presented in this Systematic Literature Review can be very useful to linked data community, since it gathers evidences from primary studies included in the review, forming a body of knowledge regarding the use of software tools to publish or consume linked data. We can outline the contribution of this work as follows:

- Research agenda based on SRL observations and according to gaps in both processes. Besides, we pointed out the need of an integration environment to simplify the process execution.

Table 5. Questions and motivation

Motivation	Question
The initial step of the publication process is badly explored by the available tools.	How to develop software tools to cope with the initial phases of the publication process of Linked Data?
None solution supports the whole process of publication of connected data.	How to provide an integrated solution to simplify the publication of Linked Data?
There is not a standard and in-depth approach to the consumption process of connected data.	How to develop and share reusable consumption patterns to facilitate the consumption of Linked Data?
None solution covers the whole process of data consumption.	How to provide an integrated solution to simplify the consumption of Linked Data?
There is need to increase the number of supported serialization formats for both publication and consumption processes.	How to provide a solution to supports RDF and non-RDF serialization formats in Linked Data Publication and Consumption process?
Low number of software tools was evaluated. Additionally, these evaluations ware not focused on support publication or consumption process.	How to validate the effectiveness of the proposed studies according to supported steps?

As future work, we intend to further investigate the software tool integration. Moreover, we intend to extend this systematic review aiming to clarify the communication interface among software tools between steps in each process.

REFERENCES

Auer, S., Bühmann, L., Dirschl, C., Erling, O., Hausenblas, M., Isele, R., & Williams, H. (2012). Managing the life-cycle of linked data with the LOD2 stack. In *The Semantic Web–ISWC 2012* (pp. 1–16). Springer Berlin Heidelberg. doi:10.1007/978-3-642-35173-0_1

Barnes, M., & Martin, R. (2002). Business data linking: An introduction. *Economic Trends*, *581*, 34–41.

Bauer, F., & Kaltenböck, M. (2011). Linked open data: The essentials. Edition mono/monochrom, Vienna.

Bizer, C., Cyganiak, R., & Heath, T. (2007). How to publish Linked Data on the Web. Retrieved September 18, 2016, from http://wifo5-03.informatik.uni-mannheim.de/bizer/HowtoPublishLinkedData.htm

Bizer, C., Heath, T., & Berners-Lee, T. (2009). Linked Data - The Story So Far. *International Journal on Semantic Web and Information Systems*, *5*(3), 1–22. doi:10.4018/jswis.2009081901

Bleiholder, J., & Naumann, F. (2009). Data fusion. *ACM Computing Surveys*, *41*(1), 1–41. doi:10.1145/1456650.1456651

Corlosquet, S., Delbru, R., Clark, T., Polleres, A., & Decker, S. (2009). Produce and Consume Linked Data with Drupal! In *Proceedings of the International Semantic Web Conference* (pp. 763-778). Springer Berlin Heidelberg. doi:10.1007/978-3-642-04930-9_48

Dermeval, D., Vilela, J., Bittencourt, I. I., Castro, J., Isotani, S., & Brito, P. (2014, September). A systematic review on the use of ontologies in requirements engineering. In *Proceedings of the 2014 Brazilian Symposium on Software Engineering (SBES)* (pp. 1-10). IEEE. doi:10.1109/SBES.2014.13

Dietze, S., Yu, H. Q., Giordano, D., Kaldoudi, E., Dovrolis, N., & Taibi, D. (2012, March). Linked Education: interlinking educational Resources and the Web of Data. In *Proceedings of the 27th annual ACM symposium on applied computing* (pp. 366-371). ACM. doi:10.1145/2245276.2245347

Ding, L., Lebo, T., Erickson, J. S., DiFranzo, D., Williams, G. T., Li, X., & Hendler, J. A. (2011). TWC LOGD: A portal for linked open government data ecosystems. *Web Semantics: Science, Services, and Agents on the World Wide Web*, *9*(3), 325–333. doi:10.1016/j.websem.2011.06.002

Heath, T., & Bizer, C. (2011). Linked data: Evolving the web into a global data space. *Synthesis lectures on the semantic web: theory and technology*, *1*(1), 1-136.

Hernandes, E., Zamboni, A., Fabbri, S., & Thommazo, A. D. (2012). Using GQM and TAM to evaluate StArt-a tool that supports Systematic Review. *CLEI Electronic Journal*, *15*(1), 3–3.

Homoceanu, S., Kalo, J. C., & Balke, W. T. (2014, June). Putting Instance Matching to the Test: Is Instance Matching Ready for Reliable Data Linking? In *Proceedings of the International Symposium on Methodologies for Intelligent Systems* (pp. 274-284). Springer International Publishing doi:10.1007/978-3-319-08326-1_28

Hyland, B., & Wood, D. (2011). The joy of data-a cookbook for publishing linked government data on the web. In *Linking government data* (pp. 3–26). Springer New York. doi:10.1007/978-1-4614-1767-5_1

Kitchenham, B., & Charters, S. (2007). Guidelines for performing systematic literature reviews in software engineering. In *Technical report, Ver. 2.3*. EBSE Technical Report. EBSE.

LaPES. (n. d.). Start - state of the art through systematic review tool. Retrieved September 04, 2016, from http://lapes.dc.ufscar.br/tools/start_tool

Moss, L., Corsar, D., & Piper, I. (2012, June). A linked data approach to assessing medical data. In *Proceedings of the 2012 25th International Symposium on Computer-Based Medical Systems (CBMS)* (pp. 1-4). IEEE. doi:10.1109/CBMS.2012.6266391

Schmachtenberg, M., Bizer, C., & Paulheim, H. (2014, October). Adoption of the linked data best practices in different topical domains. In *Proceedings of the International Semantic Web Conference* (pp. 245-260). Springer International Publishing. doi:10.1007/978-3-319-11964-9_16

Shadbolt, N., OHara, K., Berners-Lee, T., Gibbins, N., Glaser, H., Hall, W., & schraefel, . (2012). Linked open government data: Lessons from data.gov.uk. *IEEE Intelligent Systems*, *27*(3), 16–24. doi:10.1109/MIS.2012.23

Villazón-Terrazas, B., Vilches-Blázquez, L. M., Corcho, O., & Gómez-Pérez, A. (2011). Methodological guidelines for publishing government linked data. In *Linking government data* (pp. 27–49). Springer New York. doi:10.1007/978-1-4614-1767-5_2

World Wide Web Consortium. (2014). Best practices for publishing linked data. Retrieved September 04, 2016, from https://www.w3.org/TR/ld-bp/

ENDNOTES

1 https://www.w3.org/2001/sw/wiki/Category:Tool
2 http://bit.ly/2c7UlHc
3 http://bit.ly/2cB9CzE
4 http://www.w3.org/TR/ld-bp/#HTTP-URIS

Armando Barbosa is a Master's student in Informatics graduate program, Computing Institute at the Federal University Alagoas (Brazil) and a member of the Center of Excellence for Social Technologies. Armando's main research line is Semantic Web, Linked Data.

Ig Ibert Bittencourt is Professor at the Federal University of Alagoas (Brazil) and CoDirector of the Center of Excellence for Social Technologies. He received his PhD in 2009 from the Federal University of Campina Grande, Brazil and his Post-Doctoral degree in 2013 at the University of Campinas (UNICAMP). He is also the vice-president of the Special Committee of Computers and Education from Brazilian Computer Society and W3C Advisory Committee Representative. Ig Bittencourt also founded a Company in Intelligent Tutoring Systems awarded as the most innovative Brazilian company in 2014 (USP Innovation Prize). He believes in innovative social entrepreneurship as a model for promoting a sustainable economic and social development. His main interests are research, development and innovation in Computers and Education, Ontologies and Social Entrepreneurship.

Sean W.M. Siqueira is Professor at the Department of Applied Informatics, Federal University of the State of Rio de Janeiro (UNIRIO), Brazil, where he teaches courses in Information Systems, Databases, Web Science and Social & Semantic Web. He holds an MSc (1999) and a PhD (2005) in Computer Science, both from the Pontifical Catholic University of Rio de Janeiro (PUC-Rio), Brazil. His research interests include knowledge representation and management; collaborative systems; social web; semantic web; ontologies; information integration; semantic models; user models; and learning & education. He has experience in the Computer Science area, with focus on Web Science, Information Systems and Technology Enhanced Learning. He has participated in some international research projects and has written more than 130 papers for conferences, journals, and book chapters. He was the coordinator of the Graduate Program in Information Systems at UNIRIO from July 2012 to September 2014, the coordinator of the Program Committees of the Brazilian Symposium on Information Systems (SBSI 2015), the Brazilian Symposium on Computers and Education (SBIE 2012 and SBIE 2014), member of the steering committee of the Brazilian Congress of Computers and Education (CBIE 2015), co-editor-in-chief of the iSYS: Brazilian Journal of Information Systems as well as the Brazilian Journal on Computers in Education, and member of the special committee on Computers and Education (CEIE) and coordinator of the special committee on Information Systems (CESI), both from the Brazilian Computer Society.

Rafael de Amorim Silva received his BS degree in Information System from the Faculty of Alagoas (2005), Brazil. He received his MS degree in Computer Science from the Federal University of Pernambuco (2008), Brazil, and his PhD degree in Electronics Engineering from the Aeronautics Institute of Technology, Brazil. He is currently professor of the Federal University of Alagoas, Brazil. His main scientific interests include Computer Networks, Informatics in Education, and Artificial Intelligence.

Ivo Calado holds a BSc (2008) and MSc (2010) in Computer Science at Federal University of Alagoas and Federal University of Campina Grande, respectively. He also holds a PhD in Electrical Engineering at Federal University of Campina Grande, Brazil (2015). Currently, Ivo Calado is a Professor at Federal Institute of Education, Science and Technology of Alagoas, Brazil. His main research interests include Computer Networks, Internet of Things and Semantic Web.

APPENDIX A

Extraction Form

Table 6. Extraction form

#	Study Data	Description	Relevant RQ
1	Study Identifier	Unique id for the study	Study overview
2	Date of data extraction, Authors, Year, Title, Country		Study overview
3	Article Source		Study overview
4	Type of Article	Journal, Conference, workshop, book chapter	Study overview
5	Software tool name	What are the software tool name?	Study overview
6	Version	What are the software tool version?	Study overview
7	Last Update	When was the last update of the software tool?	Study overview
8	License	What is the license?	Study overview
9	Platform	Web, Mobile, Desktop	Study overview
10	Language	What are/is the used language to implement the software tool?	Study overview
11	Application Context	Industrial, Academic	Study overview
12	Purpose	Publication, Consumption	RQ1
13	Steps followed	What are the steps followed of the purpose?	RQ1
14	Serializable formats	What are the serializable formats supported by the software tool?	RQ2
15	Research Method	Controlled experiment, Case study, Use case, Conceptual proof, Benchmark, Simulation, Not applicable	RQ3

APPENDIX B
Quality Assessment

Table 7. Quality assessment

ID	References	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Total Score	Quality
S01	(Aastrand, Celebi, & Sauermann, 2010)	1	1	1	1	1	0,4	0,3	1	1	0,5	1	9,2	84%
S02	(An, Kim, Lee, & Lee, 2013)	1	1	1	1	0,5	0	0,3	1	1	1	1	8,8	80%
S03	(Dimou, De Vocht, Van Grootel, Van Campe, Latour, Mannens & Van de Walle,2014)	1	1	1	1	1	0	0,1	1	1	0	0,5	7,6	69%
S04	(Lehmann, Furche, Grasso, Ngomo, Schallhart, Sellers, & Auer, 2012)	0	0	1	0	0	0,1	0	0	1	1	1	4,1	37%
S05	(Aprosio, Giuliano, & Lavelli. 2013)	1	1	1	0	0,5	0,3	0	1	1	1	0,5	7,3	66%
S06	(Araujo, Houben, Schwabe, & Hidders, 2010)	1	1	1	1	0	0	0,4	0	1	0	0,5	5,9	54%
S07	(Assaf, Louw, Senart, Follenfant, Troncy, & Trastou, 2012)	1	1	1	1	1	0,2	0	1	1	1	1	9,2	84%
S08	(Auer, Dietzold, Lehmann, Hellmann, & Aumuelle, 2009)	1	1	1	0	1	0,4	0	1	1	1	1	8,4	76%
S09	(Augenstein, Padó, & Rudolph, 2012)	1	1	1	1	0,5	0,5	0	1	1	1	1	9	82%
S10	(Lehmann & Bühmann, 2011)	1	1	1	1	0,5	0,1	0	1	1	1	0,5	8,1	74%
S11	(Bottoni & Cerian, 2014)	1	1	1	0	0,5	0	0,4	0	0	0	0,5	4,4	40%
S12	(Chang, Xu, Zhou, Shao, Li, & Yan, 2013)	1	1	1	1	0,5	0,4	0	0	0	1	0,5	6,4	58%
S13	(Chen, Ding, Wang, Wild, Dong, Sun, & Sankaranarayanan, 2010)	1	1	1	1	0,5	0,4	0	1	1	1	0,5	8,4	76%
S14	(Becker & Bizer, 2009)	0,5	1	1	0	0,5	0	0,3	0	0	1	0,5	4,8	44%
S15	(Bizer, Lehmann, Kobilarov, Auer, Becker, Cyganiak, & Hellmann, 2009)	1	1	1	1	0,5	0,4	0	0	1	1	0,5	7,4	67%
S16	(Ciravegna, Gentile, & Zhang, 2012)	1	1	1	0	0,5	0	0,3	0	0	0	0,5	4,3	39%
S17	(Coletta, Castanier, Valduriez, Frisch, Ngo, & Bellahsene, 2012)	1	1	1	1	0,5	0,2	0	0	1	0	0,5	6,2	56%

continued on following page

Table 7. Continued

ID	References	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Total Score	Quality
S18	(Fernández, Arias, Sánchez, Fuentes-Lorenzo, & Corcho, 2014)	1	1	1	1	0,5	0,1	0	1	1	1	1	8,6	78%
S19	(Corlosquet et al., 2009)	1	1	1	1	1	0,4	0	1	1	0	0,5	7,9	72%
S20	(Tao, Song, Sharma, & Chute, 2013)	1	1	1	1	0,5	0,1	0	1	1	1	0,5	8,1	74%
S21	(Agatonovic, Cunningham, & Damljanovic, 2012)	1	0	1	1	0,5	0,1	0	1	1	0	0,5	6,1	55%
S22	(Curry, O'Donnell, Corry, Hasan, Keane, & O'Riain, 2013)	1	1	1	0	1	0,3	0	0	1	0	1	6,3	57%
S23	(Di Noia, Mirizzi, Ostuni, Romito, & Zanker, 2012)	0,5	0	0	0	0,5	0	0,3	1	1	0	0,5	3,8	34%
S24	(De Vocht, Softic, Mannens, Ebner, & Van de Walle, 2014)	1	1	1	0	1	0	0,3	1	1	1	0,5	7,8	71%
S25	(Darari & Manurung, 2011)	0,5	1	1	0	0,5	0,5	0	0	1	1	0,5	6	55%
S26	(Dastgheib, Kochut, & Mesbah, 2013)	1	1	1	1	0,5	0	0,1	0	1	1	0,5	7,1	65%
S27	(Groza, Handschuh, Möller, & Decker, 2008)	1	1	1	1	1	0,2	0	0	1	1	0,5	7,7	70%
S28	(Delbru, 2009)	1	1	1	0	1	0	0,3	1	1	0	0,5	6,8	62%
S29	(Auer, Dietzold, & Doehring, 2010)	1	1	1	1	0,5	0,1	0,3	1	1	1	1	8,9	81%
S30	(Paret, Van Woensel, Casteleyn, Signer, & De Troyer, 2011)	1	1	1	0	0,5	0	0,3	1	1	1	0,5	7,3	66%
S31	(Elze, Hesse, & Martin, 2011)	1	1	1	1	1	0	0,3	0	1	1	0,5	7,8	71%
S32	(Emaldi, Lázaro, Laiseca, & López-de-Ipiña, 2012)	0,5	1	0,5	1	0,5	0	0,3	1	1	1	0,5	7,3	66%
S33	(Tummarello, Morbidoni, Bachmann-Gmür, & Erling, 2007)	1	1	1	0	0,5	0,3	0	1	1	1	1	7,8	71%
S34	(Baritakis, Fafalios, & Tzitzikas, 2014)	1	1	1	1	0,5	0	0,3	1	1	0	0,5	7,3	66%
S35	(Farouk & Ishizuka, 2012)	1	1	1	0	0,5	0,2	0	1	1	1	0,5	7,2	65%
S36	(Fernández-Tobías, Cantador, Kaminskas, & Ricci, 2011)	0,5	1	0,5	0	0,5	0	0,3	0	1	1	0,5	5,3	48%
S37	(García, Brunetti, López-Muzás, Gimeno, & Gil, 2011)	1	1	1	1	0,5	0,1	0	0	1	1	0,5	7,1	65%

continued on following page

Table 7. Continued

ID	References	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Total Score	Quality
S38	(Aldana-Montes,García-Godoy, & Navas-Delgado, 2011)	1	1	1	1	1	0,4	0	0	1	1	0,5	7,9	72%
S39	(Graves, 2010)	1	1	1	1	0,5	0,4	0	0	1	1	0,5	7,4	67%
S40	(Wang, Liu, Penin, Fu, Zhang, Tran, & Pan, 2009)	1	1	1	1	0,5	0	0,3	1	1	1	0,5	8,3	75%
S41	(Chen, Hsueh, & Huang, 2011)	1	1	1	0	1	0,1	0	0	0	0	0,5	4,6	42%
S42	(Fisteus, García, Fernández, & Fuentes-Lorenzo, 2014)	1	1	1	1	0,5	0,1	0	1	1	1	0,5	8,1	74%
S43	(Han, Finin, Parr, Sachs, & Joshi, 2008)	0,5	1	1	0	0,5	0,4	0	1	0	1	0,5	5,9	54%
S44	(Bauer, Kaltenböck, & Recheis, 2011)	1	1	1	0	0,5	0,4	0	0	0	0	0,5	4,4	40%
S45	(Auer, Hladky, & Khalili, 2012)	1	1	1	1	0,5	0,1	0	1	1	1	0,5	8,1	74%
S46	(Auer & Khalili, 2013)	1	1	1	1	0,5	0,1	0	1	1	1	0,5	8,1	74%
S47	(Consens & Khatchadourian, 2010)	1	1	1	0	0,5	0	0,3	1	1	0	0,5	6,3	57%
S48	(Grollios, Mitkas, & Vavliakis, 2013)	1	1	1	0	0,5	0	0,1	1	1	1	0,5	7,1	65%
S49	(Konstantinou, Kouis, & Mitrou, 2014)	1	1	1	1	0,5	0,4	0	1	1	1	0,5	8,4	76%
S50	(Hernandez, Lappalainen, & Sicilia, 2013)	1	1	1	0	1	0	0,7	0	0	1	0,5	6,2	56%
S51	(Lehmann, Furche, Grasso, Ngomo, Schallhart, Sellers, & Auer, 2012)	1	1	1	0	0,5	0	0,1	1	1	0	0,5	6,1	56%
S52	(Lowe, 2009)	1	1	1	1	1	0,2	0	1	1	0	0,5	7,7	70%
S53	(Lynden, Kojima, Matono, & Tanimura, 2011)	1	1	1	1	0,5	0,2	0	1	1	1	0,5	8,2	75%
S54	(Machado & Parente de Oliveira, 2011)	1	1	1	0	0,5	0,4	0	0	1	1	1	6,9	63%
S55	(Narasimha, Kappara, Ichise, & Vyas, 2011)	1	1	1	1	0,5	0	0,3	1	1	1	0,5	8,3	75%
S56	(Curé, 2014)	1	0	1	1	1	0	0,1	0	1	1	0,5	6,6	60%
S57	(Ostuni, Di Noia, Mirizzi, Romito, & Di Sciascio, 2012)	0,5	1	1	0	0,5	0	0,3	0	1	1	0,5	5,8	53%
S58	(Otero-Garcıa, Vidal, Lama, Bugarın, & Domenech)	1	1	1	0	1	0	0,3	0	1	0	0,5	5,8	53%
S59	(Merialdo & Papotti, 2012)	1	1	1	0	0,5	0,4	0	1	1	1	0,5	7,4	67%

continued on following page

Table 7. Continued

ID	References	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Total Score	Quality
S60	(Duval & Parra Klerkx, 2010)	1	1	1	1	1	0	0,1	1	1	1	0,5	8,6	79%
S61	(Passant, 2010)	0,5	0	1	0	0,5	0	0,3	1	1	1	0,5	5,8	53%
S62	(Brambilla, Ceri, & Quarteroni, 2013)	0,5	1	1	1	0,5	0,2	0	1	1	0	0,5	6,7	61%
S63	(Radzimski, Sánchez-Cervantes, Rodríguez-González, Gómez-Berbís, & García-Crespo, 2012)	1	1	1	1	0,5	0,2	0	0	0	0	0,5	5,2	47%
S64	(Seguran, Senart, & Trastour, 2012)	1	1	1	0	1	0,1	0	0	0	0	1	5,1	46%
S65	(Shakya, Takeda, & Wuwongse, 2009)	1	1	1	1	0,5	0,1	0	0	0	1	0,5	6,1	55%
S66	(Chapman, Skarka, & Solanki, 2012)	1	1	1	0	1	0,1	0	0	1	1	0,5	6,6	60%
S67	(Solomou, Kalou, Koutsomitropoulos, & Papatheodorou, 2011)	1	1	1	1	0,5	0	0,3	1	1	1	1	8,8	80%
S68	(Sorrentino, Bergamaschi, Fusari, & Beneventano, 2013)	1	1	1	1	1	0,1	0	0	0	1	1	7,1	65%
S69	(Stadtmüller, Speiser, Harth, & Studer, 2013)	1	1	1	0	0,5	0	0,3	1	1	1	1	7,8	71%
S70	(Stolz & Hepp, 2013)	1	1	1	1	0,5	0	0,3	1	1	1	1	8,8	80%
S71	(Norheim, Roman, & Stuhr, 2011)	1	1	1	1	0,5	0,1	0	1	1	1	1	8,6	78%
S72	(d'Aquin, Motta, & Tiddi, 2014)	1	1	1	0	0,5	0	0,1	1	1	1	0,5	7,1	65%
S73	(Tramp, Heino, Auer, & Frischmuth, 2010)	1	1	1	0	0,5	0,1	0	1	1	1	0,5	7,1	65%
S74	(Tummarello, Cyganiak, Catasta, Danielczyk, Delbru, & Decker, 2010)	1	0	1	0	0,5	0	0,3	0	0	1	0,5	4,3	39%
S75	(Veres, 2012)	0,5	1	1	0	0,5	0	0,1	1	1	1	0,5	6,6	60%
S76	(Montoya, Ibáñez, Skaf-Molli, Molli, & Vidal, 2014)	1	1	1	0	0,5	0	0,3	1	1	1	0,5	7,3	66%
S77	(Gómez-Pérez, Vila-Suero, & Villazón-Terrazas, 2012)	1	1	1	1	0,5	0,4	0	1	1	1	1	8,9	81%
S78	(Zhang, Song, Li, & Liu, 2014)	1	1	1	0	0,5	0,5	0	1	1	1	0,5	7,5	68%
S79	(Kaoudi, Koubarakis, Kyzirakos, Miliaraki, Magiridou, & Papadakis-Pesaresi, 2010)	1	1	1	1	0,5	0	1	1	1	1	0,5	9	82%
S80	(Čerāns & Būmans, 2010)	1	1	1	1	0,5	0,1	0	1	1	1	0,5	8,1	74%
Average		0,92	0,92	0,97	0,55	0,61	0,14	0,13	0,62	0,83	0,74	0,61	7,08	64%