Baltic J. Modern Computing, Vol. 9 (2021), No. 4, 377-402
https://doi.org/10.22364/bjmc.2021.9.4.01

Semantic Web Technologies for Big Data Modeling from Analytics Perspective: A Systematic Literature Review

Tsvetanka GEORGIEVA-TRIFONOVA, Miroslav GALABOV

Faculty of Mathematics and Informatics, "St. Cyril and St. Methodius" University of Veliko Tarnovo, Veliko Tarnovo, Bulgaria

cv.georgieva@live.uni-vt.bg, m.galabov@ts.uni-vt.bg

Abstract. The present study addresses research on the application of semantic technologies (Semantic web technologies) to assist analysts in selecting, building, and explaining big data models. It is motivated by the established lack of a comprehensive and up-to-date systematic scientific review aimed at the use of semantic technologies for big data modeling for the purposes of their analysis. Research questions are defined, which refer to tracking the research interest in this topic; identification of the big data models to which the focus is directed and the semantic technologies applied to them and the solved analytics tasks; formulation of trends, guidelines for future work. The scientific papers included in the review are 44, collected from well-known digital libraries for scientific literature covering the period between 2011 and the beginning of 2021. As a result of the conducted research, useful conclusions are summarized for the most frequently studied big data models, semantic technologies and the research tasks solved through them.

Keywords: big data analytics, semantic technologies, systematic literature review

1. Introduction

The term big data analytics began to gain popularity after 2012, as shown in Figure 1. This figure shows a summary result from Google Trends of the number of searches in Google Search of big data analytics by years.

Big data analytics (Techopedia, 2017) is defined as an area that refers to the ways used by data scientists and various other users to analyse and systematically extract valuable information from huge volumes of data collected from a variety of sources. The approaches applied in processing must take into account the specific features of big data (Yan et al., 2020), due to which traditional business data processing systems would not be able to cope.

1.1. Big data characteristics

The main distinguishing features of big data are (Laney, 2001) 3Vs:

- Volume – the amount of generated and stored data; it is usually of the order of terabytes and petabytes;

- Velocity – the speed at which data is generated and processed, i.e. on the one hand, this characteristic refers to the data growth rate, on the other – the need for high-speed data processing and obtaining real-time results;
- Variety – providing means for processing different types of data such as structured (relational) data, semi-structured data, unstructured data.

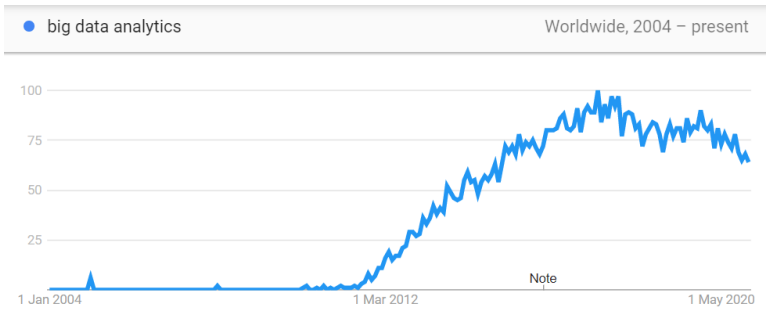

Figure 1: Trend in the use of the term big data analytics, https://trends.google.com/trends/explore?date=all&q=big%20data%20analytics

The following additional characteristics are formulated in (Seddon and Currie, 2017):

- Veracity – it is related to ensuring the trueness of the data and protecting the system from the accumulation of erroneous data. For this purpose, pre-filtering of exceptions, noise and anomalies from data sources is performed. This task is complicated by the huge volume of data and the need for high speed processing.
- Value – it is determined by the system's ability to transform data into useful information. This feature refers to the final product, the result of data processing and data analytics.
- Variability – it determines to what extent and how quickly the data structure changes, how often the meaning or format of the data varies.
- Visualization – once processed, the data must be represented in a way that is readable and accessible.

One of the promising approaches to dealing with the problems arising from the listed features of big data is the use of semantic technologies, the development of which is due to the implementation of the idea of the Semantic web.

1.2. Semantic web technologies

The ever-evolving Semantic web is a continuation of the current Web, designed to provide information for machining the semantics of large-scale data. To this end, the Semantic web provides a common framework that allows data to be represented and described so that it can be shared and reused across applications, organizations, and communities.

In terms of data, the Semantic web technologies described in (Shadbolt et al., 2006) are:

- XML (*eXtensible Markup Language*) / XML Schema are designed to define the structure of data, widely accepted for data exchange. Effective cooperation between different participants is possible only when they agree on a common syntax and have a common understanding of the basic concepts in the domain. XML covers the syntax level, but lacks support for efficient conceptual sharing.
- RDF (*Resource Description Framework*) / RDF Schema (RDFS), linked data allow semantic description of Web resources and their interrelationships in a way that is understandable to both machines and humans. RDFS allows the representation of domain knowledge using classes, properties, and instances for use in a distributed environment such as the World Wide Web.
- OWL (*Web Ontology Language*) for defining ontologies to provide a common understanding of the domain that can be shared, reused and exchanged between heterogeneous and distributed systems.
- SPARQL *(SPARQL Protocol and RDF Query Language*) for searching semantic data;
- SWRL (*Semantic Web Rule Language*) for setting rules.

1.3. Problem statement

Ensuring consistency and the ability to retrieve data for future research or reuse are the main objectives of activities and processes that are given special attention in modern flexible methodologies for implementing big data analytics such as DSE (Data Science Edge) methodology (Jurney, 2014; Grady et al., 2017). The data value pyramid proposed in (Jurney, 2014) provides a pathway from the initial data collection to the discovery of useful knowledge. Value generation increases, because the data researcher can work in the higher layers of the pyramid, after the data are refined, structured, linked, enriched with metadata and tags. The DSE methodology (Grady et al., 2017) provides a step for data curation (Singh, 2019). Such process involves performing activities such as purification, transformation, annotation in order to obtain such a representation of the data that the value of the data is preserved over time and the data remains accessible and machine-readable for reuse and storage. Therefore, the definition of an appropriate data model and their transformation and representation in the chosen model is essential as it affects the quality and efficiency of other activities such as searching, sharing, analyzing and visualizing big data. This explains the existence of research interest in the ways of modeling big data and its strengthening in the direction of the use of semantic technologies.

In the present paper, a comprehensive literature review of the scientific literature related to the application of semantic technologies for big data modeling from analytics perspective is provided. The summaries of the conducted study are the result of providing answers to a set of research questions and are represented in a form that facilitates their perception and interpretation. An analysis is made that could be useful for finding guidelines for future research and assisting the acquisition and accumulation of knowledge about the application of semantic technologies in modeling big data from the point of view of their analytics.

This paper is organized as follows. Section 2 examines existing reviews on the considered themes and identifies the need for a systematic literature review (SLR), which addresses the big data modeling through semantic technologies from analytics

perspective. Section 3 describes the research methodology. Section 4 represents and analyses the results of the review.

2. Related work

The present study is motivated by the established lack of a detailed and up-to-date systematic scientific review aimed at the use of semantic technologies to model big data for the purposes of their analytics. To this end, a study of existing scientific reviews conducted in 2015 (Ribeiro et al., 2015; Huda et al., 2015; Dou et al., 2015); 2016 (Domingue et al., 2016); 2018 (Ceravolo et al., 2018; Taouli et al., 2018); 2019 (Guedea-Noriega and García-Sánchez, 2019); 2020 (Martinez-Mosquera et al., 2020), which are discussed below.

In (Ribeiro et al., 2015; Huda et al., 2015), studies focused on big data modeling are represented. The authors identify the four main models for big data – key-value, document-oriented, width-column and graph, available when working with non-relational data. These studies demonstrate the need for data modeling as a means to improve the process of developing and analyzing big data, but they are not SLRs and do not affect the capabilities of semantic technologies in this regard.

Dou et al. (2015) examine the possibilities for supporting the semantic data mining process by exploring ways to include the formal semantics embedded in ontologies. The formal structure of the ontology allows coding of domain knowledge for data mining purposes. The authors confirm the benefit of using a formal ontology, namely a well-defined representation language, formal semantics, reasoning tools and logic inference, and consistency checking. This study is not an SLR, does not address other semantic technologies, and does not focus on the challenges posed by big data.

The study of Domingue et al. (2016) is based on a set of interviews with key stakeholders in small and large companies, and academia. Conclusions are drawn, emerging trends and future requirements for big data analytics are outlined, which include the use of semantic technologies such as RDF data, linked data, although they are reported as too complex.

Ceravolo et al. (2018) propose a literature review addressing the challenges posed by big data in data management and infrastructure. The authors acknowledge that the methods, principles, and perspectives developed by the Data Semantics community can make a significant contribution to addressing the challenges of big data. The focus of this paper is not on semantic technologies in big data modeling.

The purpose of (Taouli et al., 2018) is to explore the addition of the semantic aspect to big data analytics. A comparative study of different approaches in terms of criteria: input, output, semantics, analysis, domain, volume, diversity and speed is represented. The semantic aspect of the approaches is considered from the point of view of the three stages – big data acquisition, big data integration and big data analysis. This study is not an SLR, the comparison included 15 articles between 2009 and 2016.

In (Guedea-Noriega and García-Sánchez, 2019), a systematic review of research in the use of semantic technologies in big data analysis is proposed. The main benefits derived from the integration of semantic technologies in data analytics related to automated data processing through complex inference and reasoning techniques are highlighted; integration of heterogeneous data, data analysis at the level of knowledge;

visualization of linked data. In the present study, these advantages are explored in terms of big data modeling from analytics perspective.

Martinez-Mosquera et al. (2020) propose an SLR addressed for big data modeling and management. Big data modeling is considered at different levels of abstraction and is not aimed at the application of semantic technologies.

Table 1 contains a summary information on the similar studies discussed above – type of publication; whether it is aimed at the application of Semantic web technologies; whether it concerns big data modeling for the purposes of their analysis.

Table 1. Summary information for similar studies

Publication	**Type**	**Semantic web technologies**	**Big data**
Ribeiro et al., 2015	Survey	They are not considered.	Big data modeling and analyzing
Huda et al., 2015	Review about relational and non-relational database management systems	They are not considered.	Big data modeling
Dou et al., 2015	Survey	Ontology-based approaches; OWL	Mining big data
Domingue et al., 2016	Survey	All Semantic web technologies	Big data analyzing
Ceravolo et al., 2018	Literature review	XML, RDF, SPARQL, Linked data	Big data management in accordance with the FAIR (Findable, Accessible, Interoperable, Reusable) principles
Taouli et al., 2018	Survey	All Semantic web technologies	Semantic aspect of approaches for big data acquisition, integration, analysis
Guedea-Noriega and García-Sánchez, 2019	SLR	All Semantic web technologies	Advantages and challenges of semantic technologies in all phases of the big data analysis process
Martinez-Mosquera et al., 2020	SLR	They are not considered.	Modeling and management big data

The present study is designed as SLR. The systematic scientific review aims to answer defined research questions. For this purpose, an in-depth review of the existing literature is made on the basis of an appropriate methodology and approaches for data analysis.

Other similar types of publications are:

- Survey;

It focuses on the collection and representation of information to describe the evolution of discoveries and innovations on a given topic.

- Literature review.

It discusses the explored literature to compare different studies, draws conclusions about their weaknesses and strengths, and proposes future directions.

Publications (Guedea-Noriega and García-Sánchez, 2019) and (Martinez-Mosquera et al., 2020) are proposed as SLRs. But Martinez-Mosquera et al. (2020) does not focus on possible applications of Semantic web technologies; Guedea-Noriega and García-Sánchez (2019) raises research questions on the benefits and challenges of using semantic technologies in all phases of the big data analysis process. Therefore, open problems for SLR about the considered thematics remain the study of the built models for big data for the purpose of analysis performing; the relevant Semantic web technologies with which they are described; exploring the trends in the use of models and technologies; research problems for which they are intended.

3. Research methodology

The present study is based on the guidelines for systematic literature reviews proposed in (Kitchenham and Charters, 2007). These guidelines define the phases (planning, implementation, reporting) of the SLR process and the activities during their implementation. Following the guidelines, the activities for this SLR are carried out. Their description is exposed in section 2, where the need for such SLR is identified as part of the planning phase; sections 3 and 4, where activities from the conducing and reporting phases are represented.

3.1. Research questions

The research questions are formulated on the basis of the made study (section 2), as a result of which the need for SLR aimed at modeling big data using Semantic web technologies for the purposes of their analysis is identified.

The research questions posed in this paper are the following:

RQ1. How many research papers on the application of semantic technologies for big data modeling in order to be analyzed have been published so far? What is their distribution by years?

The aim is to establish the research interest in the considered thematics and its change over time, by providing summary information in quantitative and meaningful aspect.

RQ2. What are the big data models built for analysis and what is their distribution by years?

RQ3. Which semantic technologies are used to represent and extract big data for the purpose of performing big data analytics and what is their distribution by years?

The purpose of the research questions RQ2 and RQ3 is to identify big data models and the Semantic web technologies used for them, to explore the trends for the most popular of them.

RQ4. What research problems are addressed?

The goal is to summarize research problems that are discussed in the considered publications. In this way, it is possible to insight whether there is a predominant preference for a particular data model in solving some of the identified problems.

RQ5. What are the most frequent words and word combinations in the titles, abstracts, keywords and conclusions of the publications included in the SLR?

RQ6. Which words are most frequent occurred together in the titles, abstracts, keywords and conclusions of the publications included in the SLR?

RQ7. What are the topics derived from the titles, abstracts, keywords and conclusions of the publications included in the SLR? What data models are used for them?

RQ8. What are the most important words describing the titles, abstracts, keywords and conclusions of the publications included in the SLR in terms of data models?

For RQ5-8 research questions, a dataset (Georgieva-Trifonova and Galabov, 2021) is created, which contains the titles, abstracts, keywords, conclusions of the publications included in the present SLR. These parts of a scientific paper contain important information that includes the research problem under consideration, the proposed solution, the applied methods and approaches for this solution and can be used to support the SLR process by applying text analytics (Carnot et al., 2020), (Karami et al., 2020). Summary information is extracted from them using frequency analysis of words and word combinations, association analysis, topic modeling and feature selection. It can be used to describe the built collection of publications, as well as to support future updates of the SLR.

RQ9. What trends exist in terms of big data models, semantic technologies, and big data analytics?

The purpose of RQ9 is to find trends in the problems addressed by research in the considered themes; current issues discussed in the most recent publications, the proposed solutions for which allow future development.

The implementation of the other SLR activities corresponds to the defined research questions. The information obtained in response would be useful for researchers working in the considered and similar fields; data analysers; software engineers developing applications for big data analysis. From RQ1-4 they can acquire knowledge about the existing interest in the considered themes, as well as in the specific models, semantic technologies, problems for which they are applied; from RQ5-8 – for the built collection of publications and for the assistance of updating the made SLR; from RQ9 – for the problems to be solved in order to achieve full use of semantic technologies in modeling big data in their analysis.

3.2. Search process

The search process is a manual search in academic databases Scopus, Web of Science, EBSCO, ScienceDirect, SpringerLink, ACM Digital Library, IEEE Xplore, CiteSeerX, Google Scholar.

The search strings are obtained based on the defined research questions. Synonyms, alternative spelling, construction of more complex search strings by utilizing Boolean operators AND, OR are identified. An approach described in the methodology of (Kitchenham and Charters, 2007) is applied, in which individual aspects of the research questions are considered:

- *Population*: big data analytics, big data analysis, big data analyst, big data modeling, big data modeller, big data processing, big data curation, semantic big data, data science, data scientists;
- *Intervention*: semantic technology, Semantic Web, ontology, XML, eXtensible Markup Language, RDF, Resource Description Framework, RDFS, RDF Schema, OWL, Web Ontology Language, SPARQL, SPARQL Protocol and RDF Query Language, SWRL, Semantic Web Rule Language, linked data, triplet, triplestore, knowledge graph;
- *Comparison*: data mining, knowledge discovery, relational database, non-relational database, SQL, Structured Query Language, NoSQL;
- *Outcomes*: modeling, transforming, integration, semantic interoperability, searching, analyzing, publishing, sharing, visualization.

The search strings are constructed by concatenating the keywords from one list of each aspect with the Boolean operator OR, after which the resulting expressions are concatenated with the Boolean operator AND.

The defined inclusion and exclusion criteria are applied to the initially found publications. In addition, after the primary selection, the lists of literature sources of the selected publications are reviewed. Additionally, a search is conducted in lists of publications of authors found on their Web pages, profiles in academic social networks (such as ResearchGate, Academia.edu, Mendeley, Google Scholar).

3.3. Inclusion and exclusion criteria

The following criteria for inclusion in the review are set for the selection of the literature sources:

- Publications related to the application of semantic technologies;
- Publications in which the considered problems and semantic technologies refer to big data modeling for the purpose of their analytics;
- Publications that are papers in scientific journals, reports at scientific conferences, the full text of which is written in English, because the scientists and practitioners most often use such publications to obtain information and disseminate new discoveries.

The exclusion of literature sources from this review is based on the following exclusion criteria:

- Research reviews on the application of semantic technologies for big data modeling or analytics;
- Duplicate publications of the same study. In such cases, only the more complete version of the study is included in the review;
- Publications that are not reports at scientific conferences or papers in scientific journals (such as books, textbooks, editorial notes, dissertations, master's theses) or their full text is not written in English.

3.4. Quality assessment

After searching by keywords, 1373 publications are found. The criteria for inclusion and exclusion are applied, as well as additional research of the literature sources of the selected publications and lists of publications of authors, after which another 8 are

found. As a final result, a set of 44 publications is collected, based on which the results in section 4 are obtained.

The search process is summarized in Fig. 2 using the PRISMA Flow diagram (Moher et al., 2009).

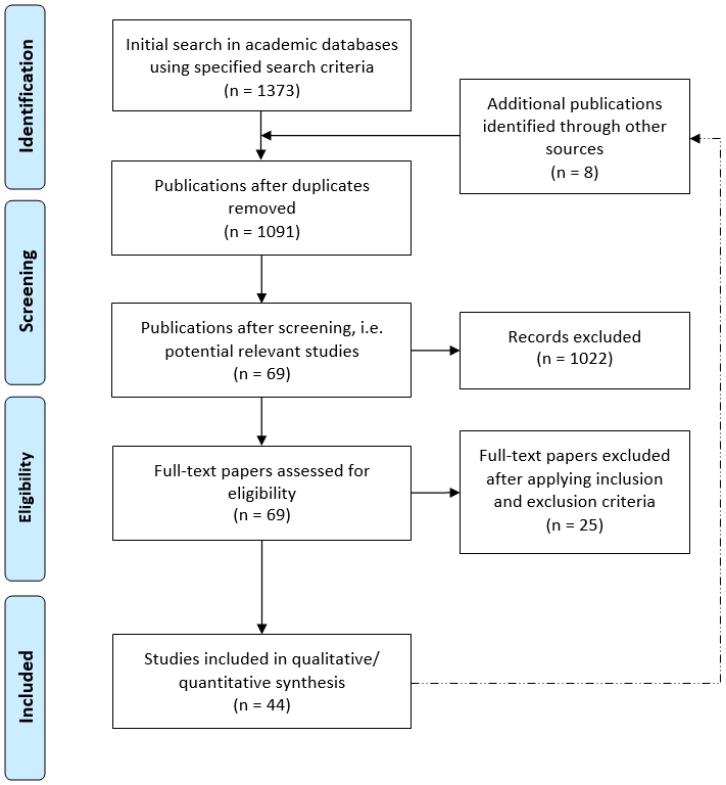

Figure 2: PRISMA Flow diagram of the publication search process for the present SLR

3.5. Data collection

The data extracted from each publication are:

- Bibliographic description – authors, title, scientific journal or conference, year of publication, as well as annotation, keywords, conclusion;
- Main thematic scope;
- Considered research problems;
- Applied or discussed semantic technologies;
- Subject area in which the results are applied (application domain);
- Datasets used in experiments;
- Guidelines for future research work.

3.6. Data analysis

The data are represented in tabular, graphical or list form to reflect the following summary information:

- Diagram of the number of publications by years (Fig. 3);
- Table for the number of publications by their type (Table 2);
- Diagram of the number of publications by data models (Fig. 4) and a diagram of the data models used in the publications by years (Fig. 5);
- Diagram of the number of publications on semantic technologies (Fig. 6) and diagram of the semantic technologies used in the publications by years (Fig. 7);
- Results of the trend exploration of the most popular data models and semantic technologies by linear regression (Table 3, Figures 8-11);
- Diagram of the research problems considered in the SLR publications and the data models used for them (Fig. 12);
- Diagram (Fig. 13) and word cloud (Fig. 14) for the frequency of occurrence of words and combinations of words in publications;
- Diagram with the words that are most frequent found together in the titles, abstracts, keywords and conclusions of the publications, together with the values of the support parameter (Fig. 15);
- Diagram with the found association rules from the words in the titles, summaries, keywords and conclusions of the publications, as well as the values of the parameters support and confidence (Fig. 16);
- Table with the words describing the extracted topics after applying an algorithm for topic modeling; the papers related to these topics (Table 4);
- Diagram of the extracted topics, the data models used for them (Fig. 17);
- Cloud with the most important words describing the titles, summaries, keywords and conclusions of the publications regarding the data models (Fig. 18).

4. Results and discussion

As a result of our search, exploration, application of the defined inclusion and exclusion criteria, 44 publications are found and selected for this review, whose research problems refer to the application of semantic technologies in big data modeling for the purpose of their analytics. The following subsections represent and discuss the results obtained in accordance with the research questions formulated in the previous section.

4.1. Distribution of publications by years

The present scientific literature review includes publications created between 2011 and the beginning of 2021. The beginning of the period can be explained by the growing popularity of the term big data analytics. The process of searching for publications and collecting data from them for the purposes of the represented review is carried out until 14 February 2021. For this reason, studies published later are not considered.

The distribution of publications by years is shown in Fig. 3.

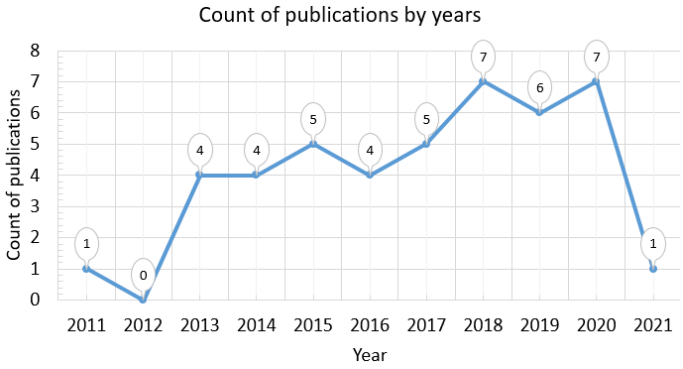

Figure 3: Diagram of the number of publications by years

From the distribution of publications by years it can be concluded that there is a tendency to increase interest in research related to the application of semantic technologies in the big data modeling in order to analyze them. As a result, the advantages of their usage are studied and practically confirmed. After defining semantic models for big data analytics in different domains and representing the existing data in their correspondence, the research interest remains relatively constant and is focused on their application to facilitate the search for useful information through scalable processing and visualization.

The most significant part of the considered publications are papers in scientific journals. The distribution of publications by type is summarized in Table 2.

Table 2. Distribution of publications by type

Type of publication	Count of publication
Proceedings paper	19
Journal paper	25

The most preferred scientific conference for representing results on the themes covered is the International Semantic Web Conference.

4.2. Data models represented by semantic technologies for the purposes of big data analytics and their distribution by years

The papers included in this scientific literature review use the following data models represented by semantic technologies for big data analytics purposes: XML graph; RDF graph; RDFS ontology; OWL ontology. Fig. 4 shows that the research interest is strongest in the OWL ontology and weakest in the XML graph.

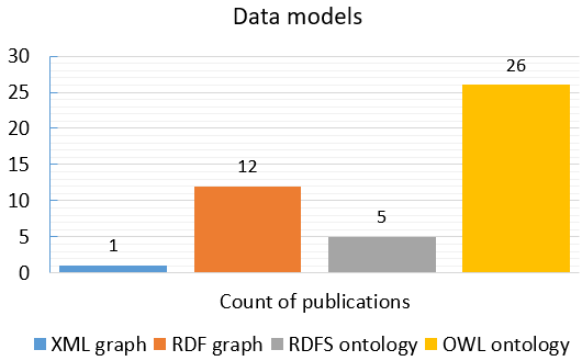

Figure 4: Diagram of the number of publications by data models

The distribution of publications by data models and by years is shown in Fig. 5. In recent years, it is noticed the presence of a growing interest in OWL ontology as a way to model big data in their analytics.

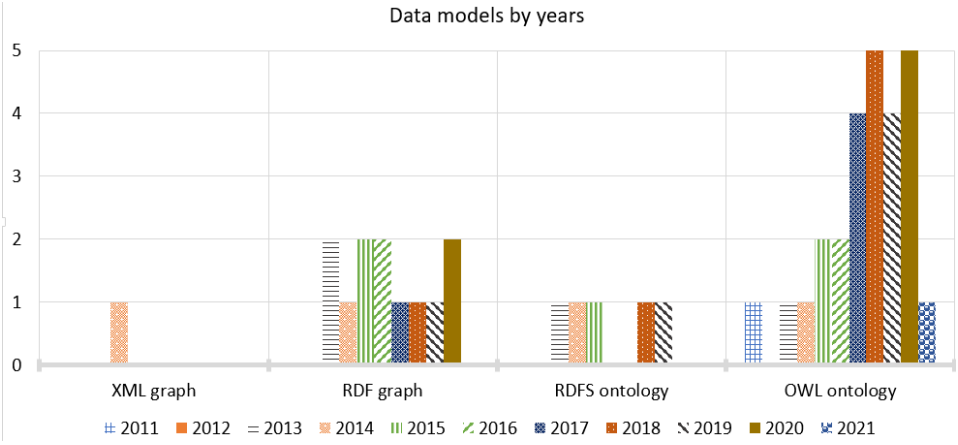

Figure 5: Diagram of the number of publications by data models and by years

4.3. Semantic technologies used in the representation and extraction of big data for the purposes of their analytics

The least used technology is XML, the most commonly used are RDF, SPARQL, OWL (Fig. 6).

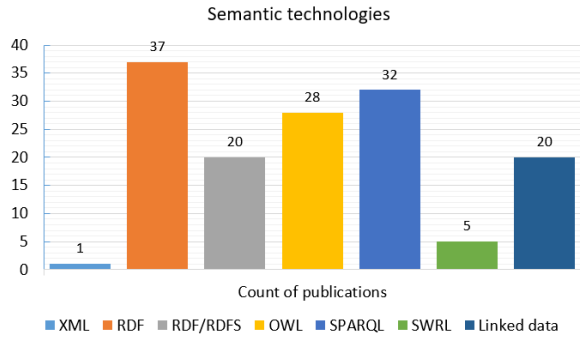

Figure 6: Diagram of the number of publications by semantic technologies

The distribution of publications by semantic technologies and by years is shown in Fig. 7.

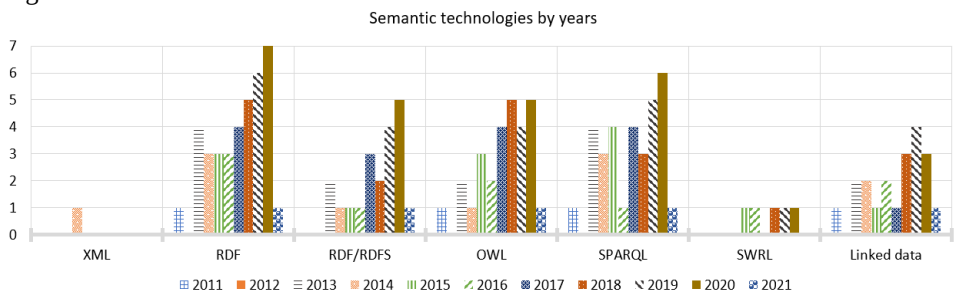

Figure 7: Diagram of the number of publications by semantic technologies and by years

A tendency to keep or increase the interest is observed for almost all semantic technologies except XML.

4.4. Trend exploration

Linear regression is applied to explore the trends of the most popular data models and semantic technologies. The process of searching for publications for the purposes of the scientific review is carried out until the beginning of 2021, therefore the study of trends includes only publications up to and including 2020. The results are summarized and visualized in Table 3 and Fig. 8-11.

Table 3. Results from applying linear regression for the publications' count by data models / semantic technologies and years (* for $p<0.1$; ** for $p<0.05$; *** for $p<0.01$)

Data model / semantic technology	Slope	*p*-value	R Square
OWL ontology	0.564	0.0001 (***)	0.859314456
RDF	0.630	0.0004 (***)	0.811281128
SPARQL	0.455	0.0189 (**)	0.518098922
OWL	0.527	0.0003 (***)	0.816240699

They show positive values of the slope of the trend line and consequently the predicted values mark a growing interest in the respective models and technologies. Statistical significance is observed for the OWL ontology; RDF, OWL (as semantic technologies) at level 0.01, as well as for SPARQL – at level 0.05.

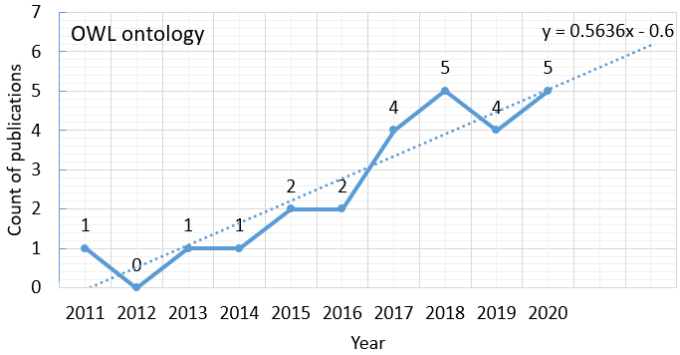

Figure 8: Applying linear regression for the count of publications that use OWL ontology as a data model by year

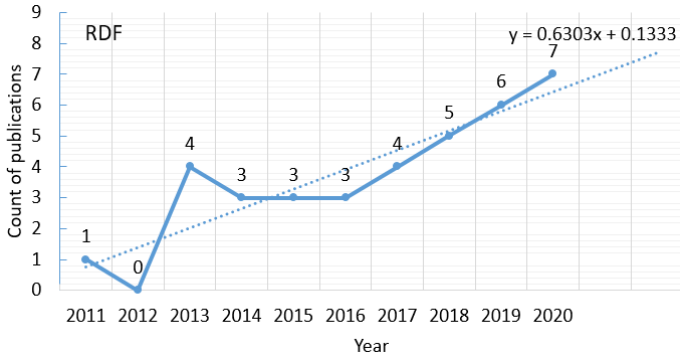

Figure 9: Applying linear regression for the count of publications that use RDF as a semantic technology by year

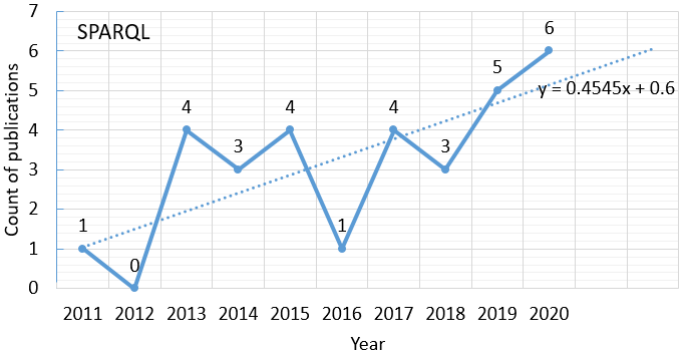

Figure 10: Applying linear regression for the count of publications that use SPARQL as a semantic technology by year

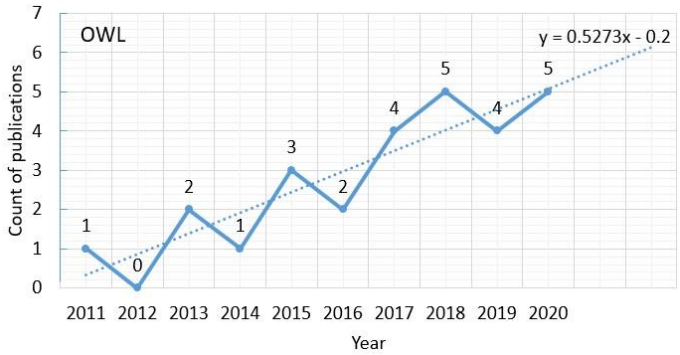

Figure 11: Applying linear regression for the count of publications that use OWL as a semantic technology by year

4.5. Considered research problems

The main research issues addressed in the papers included in this scientific literature review can be summarized as follows:

RP1. Providing easy-to-use tools for browsing, researching, analyzing, visualizing linked data that do not require in-depth knowledge of semantic technologies;

The increase in the amount of semantic data available on the Web and the insight into their potential lead to developments related to overcoming the difficulties for users to explore and use them, especially for those who have no experience with Semantic web technologies. The solutions proposed are aimed at extracting data from a dataset when its vocabulary is unknown in advance (Presutti et al., 2011); creating a framework for analysis and visualization of linked data (Klímek et al., 2013); a formal model that allows data to be linked dynamically through visualizations (Brunetti et al., 2013); query and visualization wizards (Sabol et al., 2014); real-time linked data aggregator, intuitive for biomedical experts (Kamdar et al., 2014).

RP2. Supporting data analysts and data scientists in selecting appropriate big data analysis algorithms (Nural et al., 2015; Lytvyn et al., 2018); in scalable visual examination of RDF graphs (Bikakis et al., 2016; Viola et al., 2018); in geovisual analytics (Ding et al., 2020);

RP3. Intelligent methods for organizing and processing multimedia resources (Hu et al., 2014; Rogushina et al., 2018; Greco et al., 2020);

RP4. OLAP (*OnLine Analysis Processing*) analysis for RDF data (Saad et al., 2013; Colazzo et al., 2014; Akbari-Azirani et al., 2015; Beheshti et al., 2016; Papadaki et al., 2020; Schuetz et al., 2020);

RP5. Modeling and building ontologies in different domains – healthcare (Shah et al., 2015); steel manufacturing (Bao et al., 2016); the insurance industry (Koutsomitropoulos and Kalou, 2017); drug discovery (Kanza and Frey, 2019); building waste analysis (Bilal et al., 2017); bioinformatics (Chen et al., 2020); social media (Wongthontham and Abu-Salih, 2018); regulatory reporting (Browne et al., 2019); cybersecurity (Leenen and Meyer, 2016); COVID-19 data (Kachaoui et al., 2020); the model of the learning process (Okoye, 2018); electrical utilities (Larhrib et al., 2020); Web content management systems (Vogt et al., 2019).

RP6. Integrate heterogeneous data to ensure their reuse and interoperability (Boury-Brisset, 2013; Esposito et al., 2015; Nuzzolese et al., 2017; Eine et al., 2017; Galkin et al., 2018; Karim et al., 2018; Vidal et al., 2019; Louarn et al., 2019);

RP7. Supporting the decision-making process through RDF representation of business rules (Sajjad et al., 2019);

RP8. Effective and scalable management, processing, analysis of big data represented through semantic technologies (Kim et al., 2015; Cuzzocrea et al., 2017; Belcao et al., 2021).

Belcao et al. (2021) discuss the problem of continuously increasing the volume of semantic data (i.e. modelled and represented by semantic technologies) and propose a solution for using distributed storage platforms and distributed computing machines for their processing.

Fig. 12 shows a diagram of the research problems discussed in the publications included in the present SLR and the data models used for them.

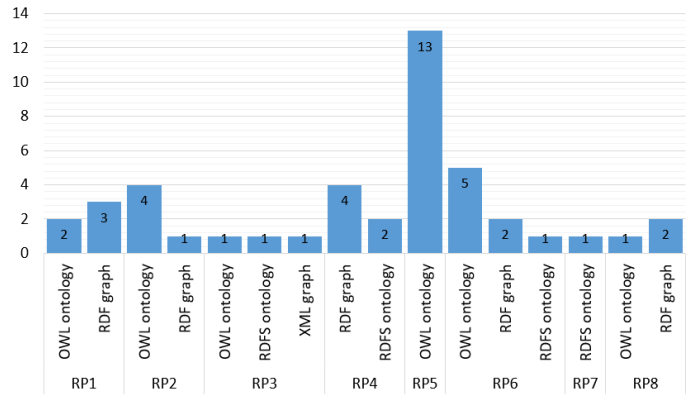

Figure 12: Diagram of the research problems discussed in the publications included in the present SLR and the data models used for them

It can be noticed that the studies related to research problem RP5 benefits from OWL over RDFS in building ontologies in various subject areas. For the rest of the research problem, there is no categorically expressed preference for a data model represented with semantic technologies. For RP2 (supporting data analysts and data scientists) and RP6 (integrate heterogeneous data), the predominant choice is OWL ontology.

4.6. The most frequent words and combinations of words in the titles, abstracts, keywords and conclusions of the publications included in the present SLR

In accordance with the justification in section 3.1, a collection of text documents consisting of the titles, abstracts, keywords and conclusions of the publications included in the present SLR is built.

The most frequent word in the resulting document collection is data. It is contained in all documents in the collection, so after tokenization it is removed together with all found stop words. The next most common words and word combinations are illustrated

in Fig. 13 and 14. The diagram in Fig. 13 shows the total number of occurrences in the entire collection and the number of documents in which the word or word combination appears at least once.

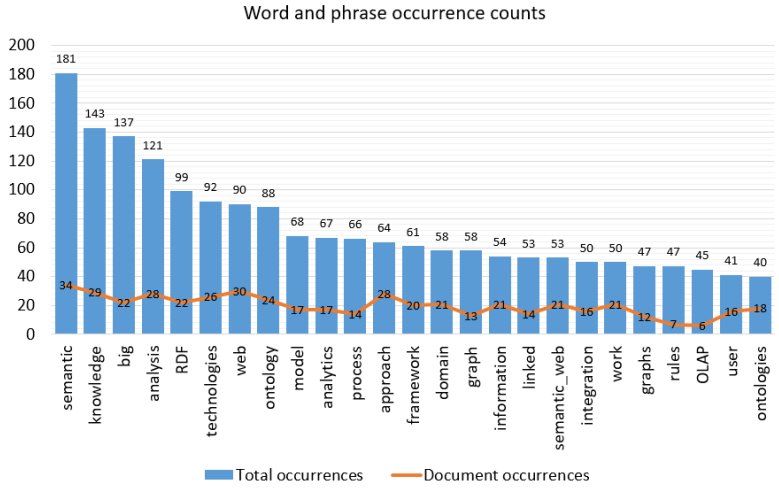

Figure 13: Diagram of the number of occurrences of words and word combinations in the publications

Figure 14: Word cloud for the number of occurrences of words and word combinations in the publications

4.7. The most frequently appearing words together in the titles, abstracts, keywords and conclusions of the publications included in the present SLR

In order to find the words that are often found together in the documents of the collection, we apply the FP-Growth algorithm for frequent patterns (Han et al., 2000). Fig. 15 shows the frequent word sets in the resulting collection of documents with a minimum support value of 35%.

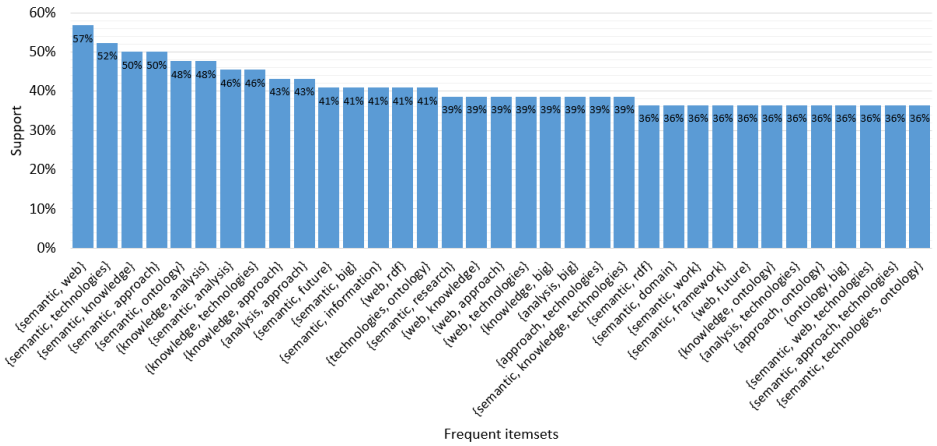

Figure 15: Diagram of the words most frequently found together in the titles, abstracts, keywords, and conclusions of publications, and the values of the support parameter

Figure 16 often shows the association rules of words in the created collection of documents with a minimum confidence value of 80%.

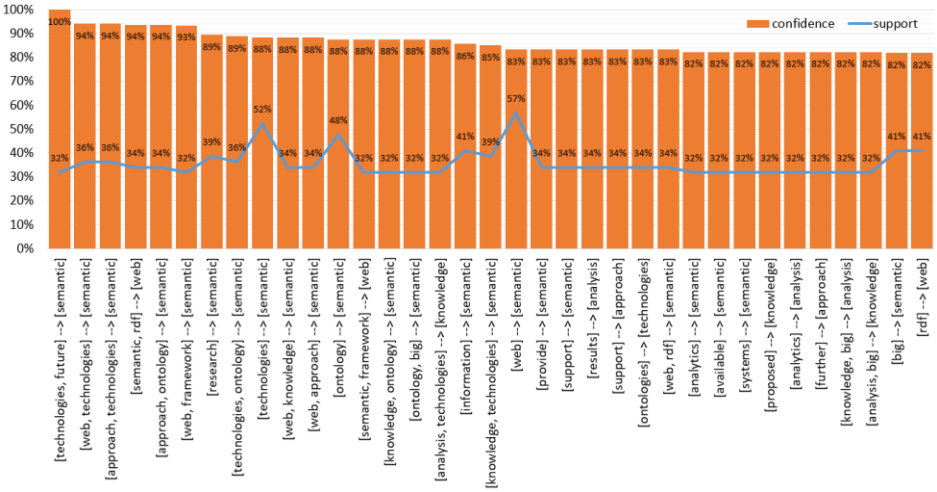

Figure 16: Diagram of the found association rules of the words in the titles, abstracts, keywords and conclusions of the publications, and the values of the parameters support and confidence

One of the frequent itemsets for documents is {*analytics*, *big*, *knowledge*} with a value of the parameter support 0.3182, consequently 31.82% of the documents in the created collection contain all these words. From this frequent itemset, the association rule {*analytics*, *big*} -> {*knowledge*} is validated with a value of the confidence parameter 0.8235, which means that 82.35% of the documents in the collection containing {*analytics*, *big*} include also the word {*knowledge*}.

4.8. Topics derived from the titles, abstracts, keywords and conclusions of the publications included in the present SLR

For a given set of documents, the purpose of topic modeling is to identify the topics that cover these documents and to group them according to the topics found. The main topic is extracted, represented by a set of words that appear in the relevant documents. The Latent Dirichlet Allocation algorithm (Yao et al., 2009) is used to detect the topics covered in the documents. The words describing the found topics are listed in Table 4, as well as the publications associated with the respective topics.

Table 4. Words describing the extracted topics after applying a topic modeling algorithm

Topic	Words	Publications
Topic 0	semantic, SOCCOMAS, WCMS, resources, web, framework, ontology, multimedia, multimedia_resources, development	Hu et al., 2014; Sabol et al., 2014; Koutsomitropoulos and Kalou, 2017; Belcao et al., 2021; Browne et al., 2019; Vogt et al., 2019
Topic 1	knowledge, integration, analysis, ontology, approach, framework, big, exploration, knowledge_graph, graph	Presutti et al., 2011; Bao et al., 2016; Nuzzolese et al., 2017; Wongthontham and Abu-Salih, 2018; Galkin et al., 2018; Vidal et al., 2019; Boury-Brisset, 2013; Ding et al., 2020
Topic 2	semantic, linked, web, semantic_web, technologies, datasets, visualization, discovery, web_technologies, domain	Klímek et al., 2013; Brunetti et al., 2013; Kanza and Frey, 2019; Kim et al., 2015; Louarn et al., 2019; Kamdar et al., 2014; Karim et al., 2018; Viola et al., 2018; Greco et al., 2020; Chen et al., 2020
Topic 3	RDF, OLAP, graph, analytics, graphs, big, process, model, operations, building	Cuzzocrea et al., 2017; Papadaki et al., 2020; Beheshti et al., 2016; Akbari-Azirani et al., 2015; Schuetz et al., 2020; Colazzo et al., 2014; Bilal et al., 2017; Bikakis et al., 2016; Esposito et al., 2015; Saad et al., 2013
Topic 4	big, analysis, technologies, semantic, rules, ontology, process, information, model, approach	Nural et al., 2015; Shah et al., 2015; Eine et al., 2017; Lytvyn et al., 2018; Rogushina et al., 2018; Sajjad et al., 2019; Larhrib et al., 2020; Okoye, 2018; Kachaoui et al., 2020; Leenen and Meyer, 2016

It can be noticed that there is some correspondence between the found topics and the research problems identified in subsection 4.5, discussed in the publications of this SLR. For example, the topic marked as *Topic 0* refers to multimedia resources (RP3, RP8); *Topic 1* – for data integration and building ontologies in different domains (RP6, RP5);

Topic 2 – for their visualization (RP1); *Topic 3* – for OLAP analysis of RDF graphs (RP4); *Topic 4* – for RDF representing business rules and supporting the selection of appropriate algorithms for the big data analytics process (RP7, RP2). Figure 17 shows the data models represented with semantic technologies that are used in the publications dealing with the respective found topics.

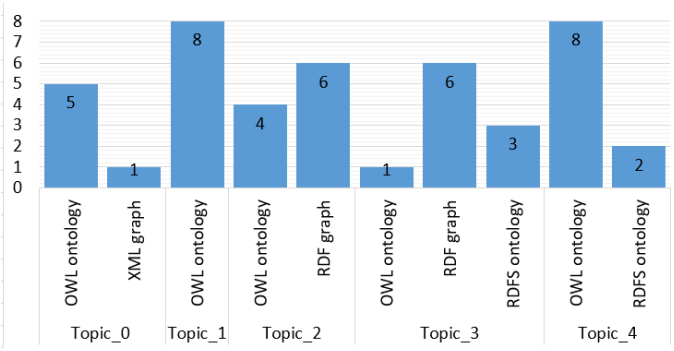

Figure 17: Diagram of the extracted topics, the data models used for them

4.9. The most important words describing the titles, abstracts, keywords and conclusions of the publications included in the present SLR in regard to data models

The feature selection for text documents consists in choosing an appropriate subset of words to increase the quality of the results of the big data analytics algorithms, as it eliminates the noise features. Different approaches to feature selection are based on the calculation and use of different importance scores or weights.

For feature selection, Gini index (Shang et al., 2007) is applied to the created collection of documents. Its purpose is to measure the impurity presence for features t_i with respect to category C_k as follows:

$$Gini(t_i) = \sum_{k=1}^{|C|} P(t_i|C_k)^2 P(C_k|t_i)^2$$

$P(t_i|C_k)$ denotes the conditional probability that a document contains the word t_i, provided that it belongs to category C_k.

Fig. 18 shows some of the most important words found, describing the titles, summaries, keywords, and conclusions of the publications regarding the data models.

The scientific research of SLR type faces challenges posed by the rapid growth of the scientific literature. It requires activities related to the extracting and processing data from a large number of scientific documents, which take a lot of time and effort and can be supported by applying text mining methods. Some of these activities are the search for new publications to update the scientific review that have emerged after its initial completion; discovery of the affected topics. Summary information on the built collection of publications, represented in subsections 4.6 – 4.9, can be used to support them. The words found as a result of frequency analysis, association analysis, topic modeling and feature selection can be used as keywords in automating the search for new future publications falling within the subject.

Figure 18: Cloud with the most important words describing the titles, summaries, keywords and conclusions of the publications regarding the data models

4.10. Observed trends in terms of big data models, semantic technologies, big data analytics

One of the observed trends is related to semantic modeling of big data, which allows their full usage as knowledge bases. On the one hand, it is related to the modeling and construction of ontologies in various domains and enrichment of existing ontologies in order to support the search and retrieval of useful information. On the other hand, there is a growing interest in integrating existing semantic data in a way that ensures their compatibility, consistency and reuse. As a result, it is possible to search for them intelligently, to inferencing, to analyze them effectively.

Another important emerging trend is aimed at filling the gap between big data and semantic technologies, related to the need for scalable management of semantic big data, which are considered as data modelled and represented through semantic technologies.

4.11. Discussion

As a result of the systematic literature review, conclusions are summarized on the current state of research related to the application of Semantic web technologies in the big data modeling for the purposes of their analysis. These conclusions can provide information on whether the use of Semantic web technologies leads to overcoming the problems associated with the big data characteristics (7Vs) mentioned in subsection 1.1; whether dealing with some issues requires compromises with others. For this purpose, in Table 5, each summary of the achievements in the considered field is compared with characteristics, the problems of which are overcome with the Semantic web technologies.

Ontologies overcome problems related to variety, veracity, value, variability, which derives from their very definition and purpose. As a result, modeling and building ontologies in different domains supports the big data analysis.

The diversity of domains is confirmed by the fact that ontologies are also used to represent information on big data analysis algorithms and their application. The

construction of such ontologies aims to assist data analysts and data scientists in selecting appropriate algorithms.

Table 5. Major contributions to the field and big data characteristics

Achievement in the field under consideration	**Big data characteristics**
Ontologies in various domains are modelled and built to support big data analytics	Variety, Veracity, Value, Variability
Enrichments of existing ontologies are described	Variety, Veracity, Value, Variability
Semantic technologies for data representation are used in order to successfully integrate heterogeneous data to ensure their reuse and interoperability	Variety, Veracity, Variability
Easy-to-use tools for browsing, exploring, analyzing, visualizing linked data are developed, which do not require in-depth knowledge of semantic technologies	Value, Visualization
Supporting the decision-making process through RDF representation of business rules is proposed	Value
Approaches for effective and scalable management, processing, analysis of big data represented through semantic technologies are implemented	Volume, Velocity

In addition, ontologies allow the development of intelligent methods for organizing and processing multimedia resources.

It is important to note that over time, enrichments of existing ontologies are needed to facilitate the search for and retrieval of useful information.

As a major challenge in big data representation with Semantic web technologies stands out achieving the scalability in the management of big semantic data to deal with the characteristics volume and velocity of big data. While the problems with the other features are "naturally" overcome by using the Semantic web technologies, in order to vanquish the problems with these two features, the existing solutions propose a combination with other modern technologies (such as Apache Spark).

5. Conclusion

The performed SLR confirms the existence of publications in which the advantages of the Semantic web technologies usage for the big data modeling for the purposes of their analysis are studied and practically validated. The main directions for future research are the full utilization of the created datasets by turning them into knowledge bases.

The comprehensive study proposed in this paper assists the acquaintance with the existing experience in the application of semantic technologies in the big data modeling in their analytics, as well as facilitates the discovery of trends and guidelines for future research.

When conducting the represented SLR, we have limited ourselves to manual search in academic databases. Our future work includes expanding the current study by performing an automatic search for publications using keywords obtained from subsections 4.6 – 4.9 for the same period of time, as well as for periodically updating it to track progress and refresh conclusions.

References

Akbari-Azirani, E., Goasdoué, F., Manolescu, I., Roatis, A. (2015) Efficient OLAP Operations for RDF Analytics, *International Workshop on Data Engineering meets the Semantic Web*, pp. 71-76.

Bao, Q., Wang, J., Cheng, J. (2016) Research on Ontology Modeling of Steel Manufacturing Process Based on Big Data Analysis, *MATEC Web of Conferences* **45**, p. 6.

Beheshti, S. M. R., Benatallah, B., Motahari-Nezhad, H. R. (2016) Scalable graph-based OLAP analytics over process execution data, *Distributed and Parallel Databases* **34**, pp. 379–423.

Belcao, M., Falzone, E., Bionda, E., Valle, E. D. (2021) Chimera: a bridge between big data and semantic technologies, *Eighteenth Extended Semantic Web Conference*, p. 16.

Bikakis, N., Liagouris, J., Krommyda, M., Papastefanatos, G., Sellis, T. (2016) graphVizdb: A Scalable Platform for Interactive Large Graph Visualization, *32nd IEEE International Conference on Data Engineering*, pp. 1342–1345.

Bilal, M., Oyedele, L.O., Munir, K., Ajayi, S.O., Akinade, O.O., Owolabi, H.A., Alaka, H.A. (2017) The application of web of data technologies in building materials information modelling for construction waste analytics, *Sustainable Materials and Technologies* **11**, pp. 28-37.

Boury-Brisset, A.-C. (2013) Managing Semantic Big Data for Intelligence, *In Proceedings of the Eighth Conference on Semantic Technologies for Intelligence, Defense, and Security*, pp. 41-47.

Browne, O., O'Reilly, P., Hutchinson, M., Krdzavac, N. B. (2019) Distributed Data and Ontologies: An Integrated Semantic Web Architecture Enabling More Efficient Data Management, *Journal of the Association for Information Science and Technology* **70**(6), pp. 575–586.

Brunetti, J. M., Auer, S., García, R., Klímek, J., Nečaský, M. (2013) Formal Linked Data Visualization Model, *In Proceedings of International Conference on Information Integration and Web-based Applications & Services* **2**(257943), pp. 309–318.

Carnot, M. L., Bernardino, J., Laranjeiro, N., Oliveira, H. G. (2020) Applying Text Analytics for Studying Research Trends in Dependability, *Entropy* **22**(11), p. 20.

Ceravolo, P., Azzini, A., Angelini, M., Catarci, T., Cudré-Mauroux, P., Damiani, E., Mazak, A., Keulen, M., Jarrar, M., Santucci, G., Sattler, K.-U., Scannapieco, M., Wimmer, M., Wrembel, R., Zaraket, F. (2018) Big Data Semantics, *Journal on Data Semantics* **7**(2), pp. 65–85.

Chen, C., Huang, H., Ross, K. E., Cowart, J. E., Arighi, C. N., Wu, C. H., Natale, D. A. (2020) Protein ontology on the semantic web for knowledge discovery, *Scientific Data* **7**(1), pp. 1-12.

Colazzo, D., Goasdoué, F., Manolescu, I., Roatis, A. (2014) RDF Analytics: Lenses over Semantic Graphs, *In Proceedings of the 23rd International conference on World Wide Web*, pp. 467–478.

Cuzzocrea, A., Buyya, R., Passanisi, V., Pilato, G. (2017) MapReduce-Based Algorithms for Managing Big RDF Graphs: State-of-the-Art Analysis, Paradigms, and Future Directions, *17th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing*, pp. 898-905.

Ding, L., Xiao, G., Calvanese, D., Meng, L. (2020) A Framework Uniting Ontology-Based Geodata Integration and Geovisual Analytics, *ISPRS International Journal of Geo-Information* **9**(8), p. 26.

Domingue, J., Lasierra, N., Fensel, A., van Kasteren, T., Strohbach, M., Thalhammer, A. (2016) Big Data Analysis. *In: Cavanillas J., Curry E., Wahlster W. (eds) New Horizons for a Data-Driven Economy*, Springer, Cham, pp 63-86.

Dou, D., Wang, H., Liu, H. (2015) Semantic Data Mining: A Survey of Ontology-based Approaches, *Proceedings of the 2015 IEEE 9th International Conference on Semantic Computing*, pp. 244-251.

Eine, B., Jurisch, M., Quint, W. (2017) Ontology-Based Big Data Management, *Systems* **5**(3), pp. 1-14.

Esposito, C., Ficco, M., Palmieri, F., Castiglionec, A. (2015) A Knowledge-Based Platform for Big Data Analytics Based on a Publish/Subscribe Service and Stream Processing, *Knowledge-Based Systems* **79**, pp. 3-17.

Galkin, M., Collarana, D., Tasnim, M., Vidal, M.-E. (2018) Synthesizing a Knowledge Graph of Data Scientist Job Offers with MINTE+, *International Semantic Web Conference*, p. 4.

Georgieva-Trifonova, T., Galabov, M. (2021) Dataset for "Semantic web technologies for big data modeling from analytics perspective: a systematic literature review, https://doi.org/10.7910/dvn/ur6own, Harvard Dataverse, V1.

Grady, N. W., Payne, J. A., Parker, H., (2017) Agile Big Data Analytics: AnalyticsOps for Data Science, *IEEE International Conference on Big Data*, Boston, MA, pp. 2331-2339.

Greco, L., Ritrovato, P., Vento, M. (2020) On the use of semantic technologies for video analysis, *Journal of Ambient Intelligence and Humanized Computing* **12**, pp. 567–587.

Guedea-Noriega, H. H., García-Sánchez, F. (2019) Semantic (Big) Data Analysis: an Extensive Literature Review, *IEEE Latin America Transactions* **17**(5), pp. 796-806. DOI: 10.1109/TLA.2019.8891948.

Han, J., Pei, J., Yin, Y. (2000) Mining frequent patterns without candidate generation, *In Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, Association for Computing Machinery, New York, NY, USA, pp. 1–12.

Hu, C., Xu, Z., Liu, Y., Mei, L., Chen, L., Luo, X. (2014) Semantic Link Network-Based Model for Organizing Multimedia Big Data, *IEEE Transactions on Emerging Topics in Computing* **2**(3), pp. 376–387.

Jurney, R. (2014) *Agile Data Science*, O'Reilly Media, Inc.

Kachaoui, J., Larioui, J., Belangour, A. (2020) Towards an Ontology Proposal Model in Data Lake for Real-time COVID-19 Cases Prevention, *International Journal of Online and Biomedical Engineering (iJOE)* **16**(9), pp. 133-136.

Kamdar, M. R., Zeginis, D., Hasnain, A., Decker, S., Deus, H. F. (2014) ReVeaLD: A user-driven domain-specific interactive search platform for biomedical research, *Journal of Biomedical Informatics* **47**, pp. 112–130.

Kanza, S., Frey, J. G. (2019) A new wave of innovation in Semantic web tools for drug discovery, *Expert Opinion on Drug Discovery* **14**(5), pp. 433-444.

Karami, A., Lundy, M., Webb, F., Dwivedi, Y. K. (2020) Twitter and Research: A Systematic Literature Review Through Text Mining, *IEEE Access* **8**, pp. 67698-67717.

Karim, R., Heinrichs, M., Gleim, L. C., Cochez, M., Porter, E., Gioia, A. L., Salahuddin S., O'Halloran M., Decker S., Beyan O. (2018) Towards a FAIR Sharing of Scientific Experiments: Improving Discoverability and Reusability of Dielectric Measurements of Biological Tissues, *In A. Paschke, A. Burger, A. Splendiani, M. S. Marshall, P. Romano, & V. Presutti (Eds.), SWAT4LS 2017: Proceedings of the 10th International Conference on Semantic Web Applications and Tools for Health Care and Life Sciences*, CEUR Workshop Proceedings, pp. 1-10.

Kim, S., Berlocher, I., Lee, T. (2015) RDF based Linked Open Data Management as a DaaS Platform, *ALLDATA 2015: The First International Conference on Big Data, Small Data, Linked Data and Open Data*, pp. 58-61.

Kitchenham, B., Charters, S. (2007), Guidelines for performing Systematic Literature Reviews in Software Engineering, *Keele University and Durham University Joint Report*, Tech. Rep. EBSE 2007-001.

Klímek, J., Helmich, J., Nečaský, M. (2013) Payola: Collaborative Linked Data Analysis and Visualization Framework. *In: Cimiano P., Fernández M., Lopez V., Schlobach S., Völker J. (eds) The Semantic Web: ESWC 2013 Satellite Events, Lecture Notes in Computer Science* **7955**, Springer, Berlin, Heidelberg, pp. 147–151.

Koutsomitropoulos, D. A., Kalou, A. K. (2017) A standards-based ontology and support for Big Data Analytics in the insurance industry, *ICT Express* **3**(2), pp. 57-61.

Laney, D. (2001) 3D Data Management: Controlling Data Volume, Velocity and Variety, *META Group Research Note* **6**(70).

Larhrib, M., Escribano, M., Cerrada, C., Escribano, J. J. (2020) Converting OCL and CGMES Rules to SHACL in Smart Grids, *IEEE Access* **8**, pp. 177255-177266.

Leenen, L., Meyer, T. (2016) Semantic Technologies and Big Data Analytics for Cyber Defence, *International Journal of Cyber Warfare and Terrorism* **6**(3), pp. 53-66.

Louarn, M., Chatonnet, F., Garnier, X., Fest, T., Siegel, A., Dameron, O. (2019) Increasing Life Science Resources Re-Usability Using Semantic Web Technologies, *15th International Conference on eScience*, pp. 217-225.

Lytvyn, V., Vysotska, V., Veres, O., Brodyak, O., Oryshchyn, O. (2018) Big Data analytics ontology, *Technology audit and production reserves* **1/2**(39), pp. 15-27.

Martinez-Mosquera, D., Navarrete, R., Lujan-Mora, S. (2020) Modeling and Management Big Data in Databases—A Systematic Literature Review, *Sustainability* **12**(2), p. 41. DOI:10.3390/su12020634.

Huda, M. M., Hayun, D. R. L., Martun, Z. (2015) Data modeling for big data, *Jurnal ULTIMA InfoSys* **6**(1), pp. 1-11.

Moher, D., Liberati, A., Tetzlaff, J., Altman, D.G. (2009) The PRISMA Group Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement, *PLoS Med* **6**(7). DOI: 10.1371/journal.pmed.1000097.

Nural, M. V., Cotterell, M. E., Peng, H., Xie, R., Ma, P., Miller, J. A. (2015) Automated Predictive Big Data Analytics Using Ontology Based Semantics, *International journal of big data* **2**(2), pp. 43–56.

Nuzzolese, A. G., Presutti, V., Gangemi, A., Peroni, S., Ciancarini, P. (2017) Aemoo: Linked Data exploration based on Knowledge Patterns, *Semantic Web* **8**(1), pp. 87–112.

Okoye, K. (2018) Mining Useful Information from Big Data Models Through Semantic-based Process Modelling and Analysis, *Communications of the IIMA* **16**(2), p. 28.

Papadaki, M.E., Tzitzikas, Y., Spyratos, N. (2020) Analytics over RDF Graphs, 13th International Workshop on Information Search, Integration, and Personalization, *Lecture Notes in Computer Science book series* **1197**, pp. 37-52.

Presutti, V., Aroyo, L., Adamou, A., Schopman B., Gangemi A., Schreiber G. (2011) Extracting core knowledge from Linked Data, *In Proceedings of the Second International Conference on Consuming Linked Data* **782**, pp. 37–48.

Ribeiro, A., Silva, A., Rodrigues da Silva A. (2015) Data Modeling and Data Analytics: A Survey from a Big Data Perspective, *Journal of Software Engineering and Applications* **8**, pp. 617-634.

Rogushina, J., Gladun, A., Pryima, S. (2018) Use of Ontologies for Metadata Records Analysis in Big Data, *In Proceedings of the XVIII International Scientific and Practical Conference "Information Technologies and Security"*, pp. 46-63.

Saad, R., Teste, O., Trojahn, C. (2013) OLAP Manipulations on RDF Data following a Constellation Model, *International Semantic Web Conference*, p. 12.

Sabol, V., Tschinkel, G., Veas, E., Hoefler, P., Mutlu, B., Granitzer, M. (2014) Discovery and Visual Analysis of Linked Data for Humans, *In: Mika P. et al. (eds) The Semantic Web, Lecture Notes in Computer Science* **8796**, Springer, Cham, pp. 309–324.

Sajjad, R., Bajwa, I. S., Kazmi, R. (2019) Handling Semantic Complexity of Big Data using Machine Learning and RDF Ontology Model, *Symmetry* **11**(309), p. 17.

Schuetz, C., Bozzato, L., Neumayr, B., Schrefl, M., Serafini, L. (2020) Knowledge Graph OLAP: A Multidimensional Model and Query Operations for Contextualized Knowledge Graphs, *Semantic Web*, pp. 1-35.

Seddon, J., Currie, W. (2017) A model for unpacking big data analytics in high-frequency trading, *Journal of Business Research* **70**, pp. 300-307

Shadbolt, N., Berners-Lee, T., Hall, W. (2006) The Semantic Web Revisited, *IEEE Intelligent Systems* **21**(3), pp. 96-101.

Shah, T., Rabhi, F., Ray, P. (2015) Investigating an ontology-based approach for Big Data analysis of inter-dependent medical and oral health conditions, *Cluster Computing* **18**(1), pp. 351–367.

Shang, W., Huang, H., Zhu, H., Lin, Y., Qu, Y., Wang, Z. (2007) A novel feature selection algorithm for text categorization, *Expert Systems with Applications* **33**(1), pp. 1-5.

Singh, D. (2019) The Role of Data Curation in Big Data, available at https://www.datasciencecentral.com/profiles/blogs/the-role-of-data-curation-in-big-data.

Taouli, A., Bensaber, D. A., Keskes, N., Bencherif, K., Badir, H. (2018) Semantic for Big Data Analysis: A survey, *International Conference on Big Data & Internet of things*, p. 16.

Techopedia (2017) Big Data Analytics, available at https://www.techopedia.com/definition/28659/big-data-analytics.

Vidal, M.E., Endris, K.M., Jozashoori, S., Karim, F., Palma, G. (2019) Semantic Data Integration of Big Biomedical Data for Supporting Personalised Medicine. *In: Alor-Hernández G., Sánchez-Cervantes J., Rodríguez-González A., Valencia-García R. (eds) Current Trends in Semantic Web Technologies: Theory and Practice. Studies in Computational Intelligence* **815**, Springer, Cham.

Viola, F., Roffia L., Antoniazzi, F., D'Elia, A., Aguzzi, C., Cinotti, T. S. (2018) Interactive 3D Exploration of RDF Graphs through Semantic Planes, *Future Internet* **10**(8), p. 30.

Vogt, L., Baum, R., Bhatty, P., Köhler, C., Meid, S., Quast, B., Grobe, P. (2019) SOCCOMAS: a FAIR web content management system that uses knowledge graphs and that is based on semantic programming, *Database: the journal of biological databases and curation* **2019**, pp.1-22. DOI: 10.1093/database/baz067.

Wongthontham, P., Abu-Salih, B. (2018) Ontology-based Approach for Identifying the Credibility Domain in Social Big Data, *Journal of Organizational Computing and Electronic Commerce* **28**(4), pp. 354-377.

Yan, Y., Zhang, L., Feng, T., Xie, P., Gao, X. (2020) Location Big Data Partition and Publishing Method based on Sampling and Adjustment, *Engineering Letters* **28**(2), pp. 280-289.

Yao, L., Mimno, D., McCallum, A. (2009) Efficient Methods for Topic Model Inference on Streaming Document Collections, *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 937–946.

Received June 21, 2021, revised August 12, 2021, accepted October 18, 2021