Heterogeneous Data Integration: A Literature Scope Review

Silvia Lucia Borowicc[1] [a] and Solange Nice Alves-Souza[2] [b]
[1]*School of Arts, Sciences and Humanities, Universidade de São Paulo, Arlindo Bettio 1000, São Paulo, Brazil*
[2]*Polytechnic School, Universidade de São Paulo, São Paulo, Brazil*

Keywords: Data Integration, Heterogeneous Data, Ontology.

Abstract: Data have been collected by communities for analysis, visualization, predictions and other activities to support data-driven decision. Obtaining value from data assets directly depends on the data integration task. However, Big Data poses new challenges to integration due to data heterogeneity. It is essential to understand the main problems and to know technologies and techniques that have been employed to improve the ability to obtain value by heterogeneous data integration. This paper presents a literature scope review that highlights the main techniques applied to heterogeneous data integration. The literature reviewed presents solutions mostly focusing on a specific purpose or part of the integration process instead of a clear understanding of how the techniques can be used in a complete integration process. Therefore, this work shows a whole picture of a data integration process organizing the techniques according to their functionalities and presents a workflow with tasks associated to techniques and resources, focusing on semantic mediation, such as mapping and matching tasks. Ontologies and semantic web technologies are promising to address data heterogeneity and have been used in the semantic enrichment of data and semantic mediation between data sources and global model. However, some aspects remain to be further investigated, such as ontology and terminology construction, data processing scalability and semantic mediation, especially for mapping definition.

1 INTRODUCTION

The existence of data, even in large volumes, is not enough to guarantee that the information demand will be effectively and timely met. Getting value from data assets in a Big Data context faces challenges related to the integration of multiple and heterogeneous data sources. The volume and heterogeneity of data hinder integration, especially when incorporating semi-structured or unstructured and semantically different data (Nathalie, 2009).

An integrated view of data can greatly contribute to obtain new information and knowledge. In the health field, for example, it is necessary to integrate several sources to assess as many risk factors as possible for a disease to manifest (Zhang et al., 2018). Likewise, in environmental analyses, it may be necessary to integrate data from different geographic dimensions or from different types of equipment and sensors to enable a complete evaluation and reduce model inconsistencies (Nundloll et al., 2021).

However, data can be spread over different organizations such as research centers, institutions and companies, which makes it difficult to cross-reference this information in computer systems. Besides technical and structural factors, the meaning of data is an aspect that increases the complexity of the integration process.

Hence, when integrating data, resources are necessary to make the data semantics explicit to allow a clear understanding of the data whose meanings are dispersed in applications and other artifacts, or even exclusively in the memory of the users. In data integration processes, metadata are essential and must be accessible for the correct use of the data.

Therefore, in Big Data contexts, to move toward solutions to integrating heterogeneous data, it is important to identify and evaluate techniques and approaches that have been employed, considering the whole data integration cycle, which includes metadata management (DAMA, 2017). Different researches rely on the semantic web, using techniques and technologies to solve semantic issues in the data integration from heterogeneous sources for specific domains (Dirgahayu et al., 2020; Kamm et al., 2021; Nagpal et al., 2021).

[a] https://orcid.org/0000-0001-7399-274X
[b] https://orcid.org/0000-0002-6112-3536

Borowicc, S. and Alves-Souza, S.
Heterogeneous Data Integration: A Literature Scope Review.
DOI: 10.5220/0012551000003690

In *Proceedings of the 26th International Conference on Enterprise Information Systems (ICEIS 2024) - Volume 1*, pages 189-200
ISBN: 978-989-758-692-7; ISSN: 2184-4992

Knowledge graphs and ontology appear in recent studies and have the potential to play an important role in data and information integration (Fathy et al., 2019; Cudré-Mauroux, 2020; Ma and Molnár, 2020), as well as data fusion with machine learning techniques, focusing on the Big Data variety(Divya and Manish, 2020; Kumar and Das, 2019).

This review highlights the requirements imposed by the ever-increasing demand for using data and the high complexity of integration processes due to data heterogeneity in multiple application contexts. Techniques and resources presented as solutions in the literature reviewed are detailed and organized according to the application functionality, helping other researchers to choose techniques and strategies for integrating heterogeneous data.

As far as it was possible to search, previous works focusing on heterogeneous data integration are directly linked to a specific application area or refer to part of the integration process (Noriega and Sanchez, 2019; Dirgahayu et al., 2020). Therefore, it is not simple to understand how to use the resources throughout a heterogeneous data integration cycle. Thus, this paper presents a unified view, developed from a broad investigation, considering the whole data integration cycle and without delimiting domains. This view shows how and in which tasks the different techniques, models, strategies, and patterns should be employed. We also suggest a data integration workflow, focused on semantic mediation tasks, which, in our view, facilitates integrated understanding of previous work.

This paper is structured as follows. Section 2 introduces basic concepts associated with heterogeneous data integration; Sections 3 and 4 present the review process and its results. Finally, Section 6 contains the conclusion and suggests future research.

2 HETEROGENEOUS DATA INTEGRATION-RELATED CONCEPTS

A variety of data integration approaches and techniques have been developed over time. Starting from approaches based on functional or relational models, with highly coupled solutions that provided a global data schema focused on structured data, until the incorporation of unstructured data with the advent of the internet (Ziegler and Dittrich, 2004).

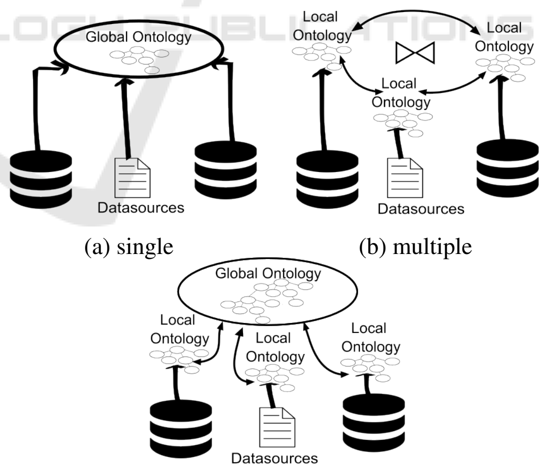
The volume and heterogeneity of data currently generated make traditional approaches difficult, especially with semantic differences (Nathalie, 2009). The syntactic and semantic heterogeneities are related to aspects such as polysemy, synonyms and abbreviations (Calva and Piedra, 2020). Ambiguities have to be eliminated when grouping, combining or completing data from different sources. Metadata, with explicit and precise semantics, can be used for correctly integrating semantically heterogeneous data.

2.1 Ontology

In information systems, ontologies specify and formalize concepts by modeling characteristics and phenomena of the world. Ontologies are defined by classes, properties, relationships and dependencies considering a specific knowledge domain (Mahmoud et al., 2021).

The use of ontologies can benefit data integration activities in several ways, according to (Xiao, 2006), including:

- metadata representation,
- automated data checking,
- global conceptualization,
- support for high-level semantic queries,
- description of the semantics of the information sources, explaining their content, and
- identification and association of semantically corresponding information concepts.



(a) single (b) multiple

(c) hybrid

Figure 1: Architectures for using ontologies to explain content.

Figure 1 shows architectures for using ontologies in heterogeneous data integration: [single] in which the data sources are mapped to a general representation model of the knowledge domain; [multiple] models from the original data sources that undergo a mapping process among themselves, that is, the global model is the result of a Cartesian product obtained by cross-referencing all the original models; and [hybrid] which uses the data source models built on a shared global vocabulary of basic terms that correspond to a domain, which allows them to be shared with each other (Wache et al., 2001).

Different approaches can be used to link ontologies to data sources, either by a relationship with the database schema or with terms in the database content. The approaches described by (Wache et al., 2001) may include the following strategies, either alone or any combination thereof:

- **Structure Resemblance:** integration occurs by generating a model that reflects the original structure in a one-to-one mapping from the ontology definitions;
- **Definition of Terms:** the ontology is used to define database or schema terms;
- **Structure Enrichment:** a combination of the two previous approaches.
- **Meta-Annotation:** meta-annotation is used to add semantic information to an information source.

Language dependency of ontology needs to be considered when addressing ontology sharing, merging, and translation, topics that often involve multiple vocabularies and conceptualizations (Guarino, 1998). To relate different ontologies, mediating agents perform the translation between the ontologies, either by lexical relationships that allow comparing language terms, or using a general ontology related to the other ontologies, or even by the search for correspondence semantics between concepts of different ontologies (Wache et al., 2001).

2.2 Semantic Web and Knowledge Graph

The interconnected datasets on the Web are called Linked Data[1] (LD). W3C refers to the network of interconnected data as the Semantic Web[2]. The technologies associated with the Semantic Web allow creating databases on the Web, as well as developing resources to interconnect and consume these data. The standard model for representing information on the Web is based on the Resource Description Framework (RDF)[3]. This model derives the link structure of the Web using URIs to identify resources (entities, concepts, objects) and relationships (interactions, events). The basic unit of data representation is a triple: ”subject, predicate, object”.

Graphs have become one of the main data structures used in heterogeneous data integration. Machine-readable and human-understandable, they can represent objects and interactions in a flexible modeling that allows mapping most types of data (Gomes and Santanchè, 2015; Jie et al., 2021).

In the semantic web, ontologies are representations based on RDF triples, and similarly, real-world entities and relationships can be represented using knowledge graphs. Data sources with their own semantic models and ontologies can be embedded in knowledge graphs (Pomp et al., 2017; Hao et al., 2021; Zhao et al., 2021).

3 LITERATURE SCOPE REVIEW PROTOCOL

The protocol adopted, based on the (Kitchenham and Charters, 2007) proposal for systematic reviews in Software Engineering, is divided into three phases: planning, conduction and results.

As shown in Figure 2, the review protocol started from an exploratory analysis of the literature to understand the concepts related to the integration of heterogeneous data. This analysis was the basis of the objective: to broadly investigate the techniques and approaches for heterogeneous data integration, considering the whole data integration cycle.

Table 1: References that supported the exploratory review.

Subject	References
data integration or heterogeneous data	(Sugawara and Nikaido, 2014);(Özsu and Valduriez, 2020);(Ziegler and Dittrich, 2004);(Batini et al., 1986);(Zhang et al., 2018)
knowledge graph	(Hao et al., 2021);(Zhao et al., 2021)
ontology	(Gruber, 1993);(Guarino, 1998);(Nathalie, 2009);(Wache et al., 2001);(Zhao et al., 2021);(Zhang et al., 2018)

Also based on the exploratory analysis, the research questions and the planning of this review were specified. Table 1 shows the references that supported the exploratory review, classified by subjects related to data integration. Reading began with reference publications and texts addressing data integration techniques to understand related terms.

[1] https://www.w3.org/standards/semanticweb/data
[2] https://www.w3.org/standards/semanticweb/
[3] http://www.w3.org/RDF/

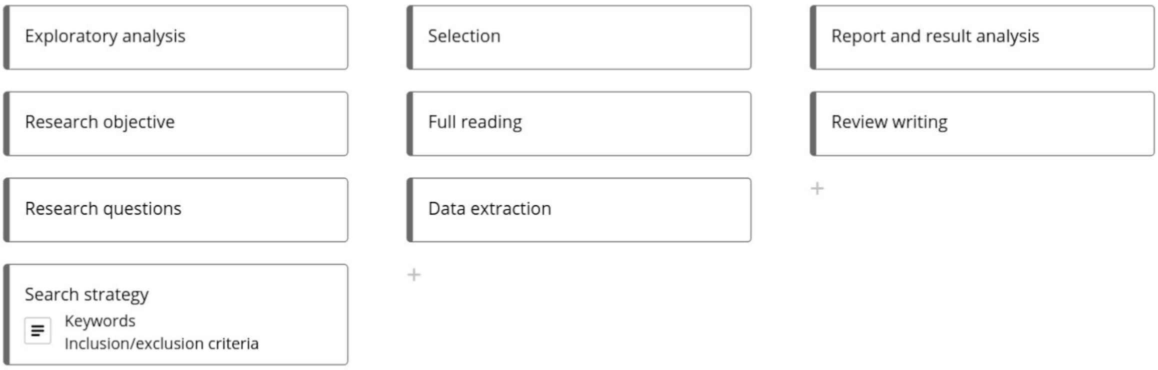


Figure 2: Literature scope review protocol.

3.1 Research Questions

Research questions (RQ1 e RQ2) are designed to explore the latest research results in heterogeneous data integration considering that different techniques and approaches can be employed in the process.
RQ1. What techniques have been used for heterogeneous data integration?
RQ2. How to bring in integrated databases from heterogeneous sources?

These questions are intended to provide an overview of the techniques used to integrate data with heterogeneous structures, syntaxes or semantics. Design and validation of frameworks, the use or implementation of tools and approaches used to meet demands for integrated data were considered. Techniques for materialized or virtual data integration were verified. The search for articles was limited to the period from 2015 to 2022.

3.2 Search Strategy

The search string comprised three keywords and their related terms, according to the exploratory analysis readings, presented in Table 2.

Table 2: Search keywords.

Keyword	Related terms
data integration	information integration
heterogeneous data	heterogeneous datasets heterogeneous datasources heterogeneous information unstructured data
ontology	knowledge graph semantic

The inclusion and exclusion criteria are respectively presented in Tables 3 and 4.

Scopus, ACM, *Engineering Village*, IEEE, *Web of Science* and PubMed were selected to search papers due to their credibility, adequacy to the computing area and health besides being paid by the university,

Table 3: Inclusion criteria (IC).

Criteria	Description
IC-1	The research addresses, applies and discusses the results of the application of heterogeneous data integration techniques.
IC-2	The research addresses, applies and discusses the results of applying integrated data modeling techniques from heterogeneous data sources.
IC-3	Paper published in a journal or conference between 01/01/2015 and 31/12/2022.

Table 4: Exclusion criteria (EC).

Criteria	Description
EC-1	Literature review
EC-2	Paper is not written in Spanish, English or Portuguese

allowing full access to their content. The search was conducted in January 2023 and the results were imported into the Parsif.al[4] tool. After excluding duplicates, 508 articles remained for selection, in the Conduction phase (Figure 3).

3.3 Conduction

A prior selection was made by reading the title, abstract and keywords. To support the selection process, ASReview[5] was used, a tool that implements active learning, with its standard classification algorithm.

ASReview reorders unread texts as it ”learns” from those that have already been annotated as relevant or irrelevant. The tool contributed to increase confidence by corroborating the results of the manual process. After selection, 81 articles were included, about 16% of the 508 articles found.

[4]https://parsif.al/
[5]https://asreview.nl/lab/

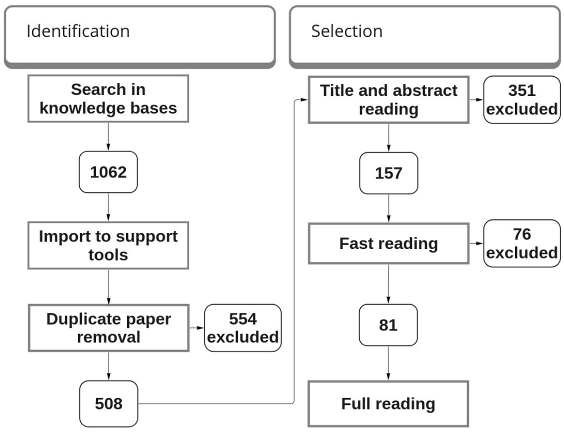


Figure 3: Conduction workflow.

4 RESULTS OF THE LITERATURE SCOPE REVIEW

The authors mainly focus on integration based on semantic web technologies. Data are mostly represented in RDF graphs and ontologies are used to organize and formalize semantic knowledge. The research on data integration and the semantic web complement each other and are often combined to solve problems of semantic data heterogeneity, promoting data sharing and efficient use from different autonomous sources (Kessler et al., 2021; Jeong and Jeong, 2015).

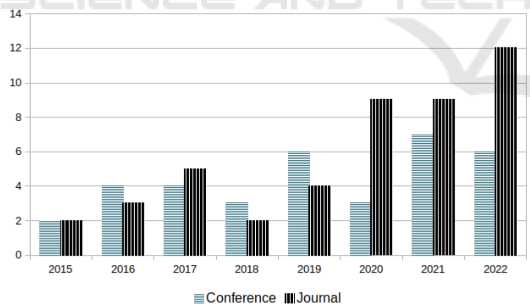


Figure 4: Papers by document type and publication year.

Figure 4 shows the distribution of papers by publication type (conference or journal) and year. The topic is observed to have had increasing interest; in contrast, conference publications decreased in 2020, at the peak of the COVID-19 pandemic, which undoubtedly affected not only attendance, but also led to the suspension of several conferences.

4.1 Semantic Enrichment in Heterogeneous Data Integration

Due to the large volume, variety, velocity and complexity of data, semantic enrichment of data and metadata for understanding the context is essential for obtaining relevant information and knowledge. Researchers have used resources such as ontologies, knowledge graphs and classification to semantically enrich data and increase the efficiency of integration processes by automating entity mapping. Table 5 lists publications according to the resources used.

Ontologies are commonly expressed formally using semantic markup languages such as RDFs or Web Ontology Language (OWL) and thus describe domains of knowledge from classes, properties and relationships. These representations are used in semantic mapping processes with integration purposes. Thesaurus and standards created for data exchange can also be used in the ontology and graph construction or in classification processes.

4.2 Global Modeling and Semantic Mediation

Regarding semantic mapping, as well as the architectures for using ontologies, in the content explication presented in Figure 1, three architectures were observed in the papers. Table 6 lists the publications according to the mapping approach presented.

In the hybrid approach used by (Nundloll et al., 2021), a different nomenclature appears to identify the global and local models. The global ontology is called domain ontology, whose level of abstraction is independent of implementations and reflects a domain of knowledge. Local ontology, called data ontology, models a specific data source and interfaces between the data and the domain ontology.

In this review, data source mapping and matching tasks are also referred to as semantic mediation. To implement data source mapping, tools are used, mainly based on RDF Mapping Language (RML)[6]. These tools are presented in Table 7 together with the formats accepted.

Among the cited tools, Karma is pointed out by (Zhang et al., 2021) as the state of the art for semantic annotation of structured data and publication in Linked Open Data. Compared to Karma, according to (Masmoudi et al., 2019), DISERTO requires less human intervention in the process by performing automated mapping but has fewer input formats allowed.

[6] https://rml.io/

Table 5: Semantic enrichment resources.

Technique	Publications
Ontology	(Gil et al., 2021);(Dridi et al., 2020);(Ding et al., 2020);(Zhang et al., 2021);(Nundloll et al., 2021);(Sernadela and Oliveira, 2017);(Rouces et al., 2018);(Fusco and Aversano, 2020);(Wang et al., 2017);(Zhang et al., 2018);(Grasso et al., 2015);(Saber et al., 2018);(Hao et al., 2021);(Masmoudi et al., 2019);(Sima et al., 2019);(Baazaoui-Zghal, 2016);(Calva and Piedra, 2020);(Masseroli et al., 2016);(Radaoui et al., 2019);(Kim et al., 2021);(Avila et al., 2019);(Rani et al., 2019);(Buron et al., 2020);(Nimmagadda et al., 2019);(Yadav et al., 2021);(SCHIESSL and BRÄSCHER, 2017);(Niang et al., 2016);(Kessler et al., 2021);(Lembo and Scafoglieri, 2020);(Jeong and Jeong, 2015);(Buron et al., 2020);(Sengloiluean and Khuntong, 2020);(Shen et al., 2016);(Cheung et al., 2015);(Xiao et al., 2017);(Zhou, 2016);(Yun et al., 2019);(Qundus et al., 2021);(Pereira et al., 2020);(Capodieci et al., 2016);(Dao et al., 2021);(Mountasser et al., 2021);(Nath et al., 2017);(Santipantakis et al., 2020);(Mrhar et al., 2020);(Mami et al., 2019);(Mahmoud et al., 2021);(Khnaisser et al., 2022);(Burgdorf et al., 2022);(Maga-Nteve et al., 2022);(Wu et al., 2022);(Ramzy et al., 2022);(Ma and Molnár, 2022);(Haghgoo et al., 2022);(Phengsuwan et al., 2022);(Katrandzhiev et al., 2022);(Krataithong et al., 2022);(Bonte et al., 2022);(Guedea-Noriega and García-Sánchez, 2022);(Thirumahal et al., 2022);(Stroganov et al., 2022)
Knowledge graph	(Nashipudimath et al., 2020);(VandanaKolisetty and Rajput, 2021);(Gupta and Gupta, 2021);(Yafooz et al., 2018);(Dhayne et al., 2018);(Bartusiak and Lässig, 2016);(Asprino et al., 2023);(Zhao et al., 2022);(Oo et al., 2022)
Classification	(Vilches-Blázquez and Saavedra, 2022);(Ma et al., 2017);(Balachandran et al., 2019);(Zhao et al., 2021);(Le Guillarme et al., 2021);(Sandhya and Roy, 2016);(Pomp et al., 2017);(Jie et al., 2021);(Vidal et al., 2019);(Gomes and Santanchè, 2015)

Moreover, (Mrhar et al., 2020) showed improvement in semantic recognition accuracy, in semi-automatic mapping, by associating Karma with an algorithm that combines Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997) with Conditional Random Field (CRF) (Lafferty et al., 2001).

Table 6: Semantic mapping architecture.

Mapping	Publications
Single	(Dridi et al., 2020);(Ding et al., 2020);(Ma et al., 2017);(Sernadela and Oliveira, 2017);(Rouces et al., 2018);(Nashipudimath et al., 2020);(Wang et al., 2017);(Saber et al., 2018);(Le Guillarme et al., 2021);(Hao et al., 2021);(Masmoudi et al., 2019);(Pomp et al., 2017);(Baazaoui-Zghal, 2016);(Calva and Piedra, 2020);(Masseroli et al., 2016);(Radaoui et al., 2019);(Gupta and Gupta, 2021);(Avila et al., 2019);(Nimmagadda et al., 2019);(SCHIESSL and BRÄSCHER, 2017);(Jeong and Jeong, 2015);(Qundus et al., 2021);(Capodieci et al., 2016);(Santipantakis et al., 2020);(Gomes and Santanchè, 2015);(Mami et al., 2019);(Mahmoud et al., 2021);(Asprino et al., 2023);(Khnaisser et al., 2022);(Burgdorf et al., 2022);(Zhao et al., 2022);(Wu et al., 2022);(Ramzy et al., 2022);(Oo et al., 2022);(Phengsuwan et al., 2022);(Krataithong et al., 2022);(Bonte et al., 2022);(Guedea-Noriega and García-Sánchez, 2022)
Multiple	(Balachandran et al., 2019);(Grasso et al., 2015);(Sima et al., 2019);(Shen et al., 2016);(Cheung et al., 2015);(Zhou, 2016);(Vidal et al., 2019);(Dao et al., 2021);(Bartusiak and Lässig, 2016);(Nath et al., 2017)
Hybrid	(Zhang et al., 2021);(Vilches-Blázquez and Saavedra, 2022);(Nundloll et al., 2021);(Fusco and Aversano, 2020);(Zhang et al., 2018);(Sandhya and Roy, 2016);(Kim et al., 2021);(Rani et al., 2019);(Buron et al., 2020);(Niang et al., 2016);(Kessler et al., 2021);(Buron et al., 2020);(Sengloiluean and Khuntong, 2020);(Xiao et al., 2017);(Yun et al., 2019);(Dhayne et al., 2018);(Mountasser et al., 2021);(Mrhar et al., 2020);(Maga-Nteve et al., 2022);(Thirumahal et al., 2022)

Table 7: Data source mapping tools

Tool	Format	Publications
DISERTO	CSV; ENVI	(Masmoudi et al., 2019)
HL7toRDF	HL7	(Dhayne et al., 2018)
Karma	DB; DSV; XML; JSON; KML	(Zhang et al., 2021)(Xiao et al., 2017)(Yun et al., 2019)(Qundus et al., 2021)(Capodieci et al., 2016)(Mrhar et al., 2020)
OAM	DB	(Krataithong et al., 2022)
Ontop	ERDB	(Niang et al., 2016)(Sima et al., 2019)(Kessler et al., 2021)(Ding et al., 2020)(Zhang et al., 2018)

As regards performing semantic queries on data, the predominant language used to create and process queries is SPARQL, which is a W3C standard language capable of retrieving and manipulating data represented in RDF (Dao et al., 2021).

Traditional architectures widely used for data integration, such as Data Warehouse (DW), can also implement semantic enrichment to solve heterogeneity problems. Ontology is used in metadata that semantically describes models or in defining relational or multidimensional database schema.

Table 8 presents some detailed publications implementing the link between the semantic model and data. Data sources and different approaches for representing the semantic model are presented. The approaches are related to the linking strategies presented in Section 2.1.

Table 8: Connection between semantic model and data sources.

Source	Connection	Publications
DW	Definition of Terms: schema	(Masseroli et al., 2016);(Baazaoui-Zghal, 2016)
Relational Database	Definition of Terms: schema	(Khnaisser et al., 2022)
DW	Meta-Annotation	(Nimmagadda et al., 2019)
DW	Definition of Terms: RDF	(Mahmoud et al., 2021);(Nath et al., 2017)
Knowledge Graph	Definition of Terms: RDF	(Pomp et al., 2017)

Besides traditional data structures, the graph-based data model has been considered a better choice and has become one of the main ways to unify heterogeneous data, regarding the ease of mapping most types of data due to modeling flexibility (Gomes and Santanchè, 2015; Jie et al., 2021).

By associating graphs with ontologies to deal with semantic heterogeneity, it is possible to interconnect and manipulate data, building a coherent and integrated view from multiple and heterogeneous sources. The RDF graph can be interpreted using an ontology and also supports queries on data and ontology at the same time (Jie et al., 2021; Buron et al., 2020; Vilches-Blázquez and Saavedra, 2022).

A semantic model can also be created directly from the embedded data. A knowledge graph can be generated from reading the content of data sources. The semantic model is not static, being expanded as new data sources are included by users during the integration process (Pomp et al., 2017).

4.3 Data Processing Techniques

Pattern recognition and natural language processing (NLP) have been used in data integration from structured and unstructured sources, for text reading, semantic mapping, data fusion, linkage, entities recognition and classification.

NLP is used to read data from text files, PDF files or from text fields stored in relational databases. The papers presented semi-automated entity recognition processes, with user participation in data validation and cleaning. In these processes, some methods are used, such as most frequent terms index and similarity metrics. To deal with imprecise information and linguistic ambiguity fuzzy logic is used for mapping and semantic enrichment processes (Baazaoui-Zghal, 2016; Haghgoo et al., 2022; Krataithong et al., 2022; Stroganov et al., 2022) .

Large databases, ontologies, knowledge graphs in English language are predominant; nonetheless, multilingual solutions were found; a publication in which the authors (SCHIESSL and BRÄSCHER, 2017) use a database in Portuguese, and (Guedea-Noriega and García-Sánchez, 2022) worked with Spanish texts.

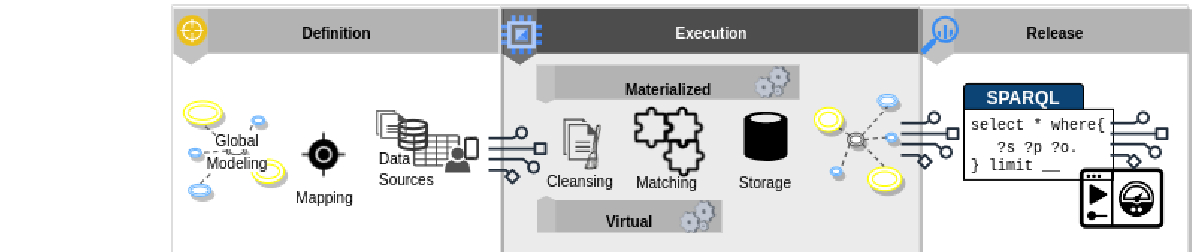
4.4 Data Storage Approaches and Scalability

The storage approaches in data integration are twofold: (i) materialized data integration built in the storage layer, i.e., data are loaded, transformed and stored in an integrated database; (ii) virtual data integration, which occurs in the query layer and whose searches are performed into a global model, although the data remains stored only in its original source.

In Big Data environments with a materialized approach, traditional resources such as DW have not been highly scalable and add cost to the storage infrastructure (Rani et al., 2019). In virtual integration approaches, global query models and federated databases are adopted to avoid materialization costs and increase scalability (Masmoudi et al., 2019). Nevertheless, in this case, there is an impact and an increase in processing infrastructure costs, since all data is transferred and processed online at the time of the request.

Big Data technologies are used in some recent works, aiming at scalability and cost reduction using distributed infrastructure with low computational power, scaling by the distribution of processing. Thereby, tools and techniques such as Hadoop and MapReduce are used to implement distributed processing in the materialized or virtual approach, in queries and mappings (Rani et al., 2019; VandanaKolisetty and Rajput, 2021; Santipantakis et al., 2020), or storage (Nashipudimath et al., 2020).

To improve performance in semantic mapping and query processing, (Rani et al., 2019; Kim et al., 2021) implemented architectures with multiple semantic levels, supporting the process of combining



	Definition	Execution	Release
Pattern Recognition	✓	✓	
Drivers	✓	✓	✓
RDF Graph	✓	✓	✓
R2ML; RML	✓		
RDFs; OWL	✓		
NLP	✓	✓	✓
SPARQL			✓

Figure 5: Resources and techniques applied to integration workflow.

ontologies and queries that can be executed at multiple levels.

5 DISCUSSION

In the articles included in the review, it was not possible to identify a workflow that demonstrated the use of techniques and resources in the stages of a data integration process. Therefore, to facilitate understanding, Figure 5 organizes knowledge, including the most common techniques found in the review, indicating at which stages they are used.

The workflow is divided into three stages: (i) definition, whereby the source and target terminologies are identified and mappings are specified; (ii) execution is the stage in which ingestion, processing and matching are performed differently depending on the data storage approach. In virtual scenarios, middleware are commonly used to split and translate federated queries to the source query language. In materialization scenarios, integrated data is transformed into RDF graphs and stored; (iii) in the release stage integrated data are available to explore, predominantly by SPARQL endpoints.

Most of the selected publications, 79 out of the 81, mention the use of some semantic enrichment resource, such as ontologies and knowledge graphs. In semi or unstructured data, natural language processing (NLP) was used to extract data and then subject the data to pattern recognition to identify entities and match them with global data models.

In papers regarding transformation tools used to perform matching according to mapping, in RML, and conversion to RDF data, data sources and RML mapping files are placed as inputs to the matching process. No mentions were found about creating this mapping in any other way other than using a tool that reads the structure of the data source and the model, submitting the mapping indication to the user or, manually writing the rules in RML, which makes knowledge of the RML language imperative to implement data transformation to RDF.

6 CONCLUSION

Heterogeneous data integration is essential to obtain value from data assets. In this review, techniques and approaches used for integrating heterogeneous data were investigated.

The review results show that ontologies and semantic web technologies are promising to resolve data heterogeneities. Also, Big Data technologies have been used in some proposals for distributed storage and query processing, or mappings, contributing to scalability. However, there are some aspects of the research, including ontology construction and semantic mediation, that remain open. Furthermore, aspects of data governance in the data integration workflow, establishing patterns focusing on semantic mediation, also remain open for further investigation.

REFERENCES

Asprino, L., Daga, E., Gangemi, A., and Mulholland, P. (2023). Knowledge graph construction with a façade: A unified method to access heterogeneous data sources on the web. *ACM Transactions on Internet Technology*, 23(1):1–31.

Avila, C., Calixto, A., Rolim, T., Franco, W., Venceslau, A., Vidal, V., Pequeno, V., and Moura, F. F. D. (2019). Medibot: An ontology based chatbot for portuguese speakers drug's users. In M., B. A. H. S. F. J. S., editor, *Proceedings of the 21st International Conference on Enterprise Information Systems*, volume 1, pages 25–36, Heraklion, Crete, Greece. SCITEPRESS - Science and Technology Publications.

Baazaoui-Zghal, H. (2016). Fuzzy ontology-based spatial data warehouse for context-aware search and recommendation. In *Proceedings of the 11th International Joint Conference on Software Technologies*, pages 161–166, Lisbon, Portugal. SCITEPRESS - Science and Technology Publications.

Balachandran, S., Ranganathan, V., and Vetriveeran, D. (2019). Aligning large biomedical ontologies for semantic interoperability using graph partitioning. *International Journal of Biomedical Engineering and Technology*, 31(2):137–160.

Bartusiak, A. and Lässig, J. (2016). Semantic processing for the conversion of unstructured documents into structured information in the enterprise context. In *Proceedings of the 12th International Conference on Semantic Systems*, SEMANTiCS 2016, pages 125–128, New York, NY, USA. Association for Computing Machinery.

Batini, C., Lenzerini, M., and Navathe, S. B. (1986). A comparative analysis of methodologies for database schema integration. *ACM Computing Surveys*, 18(4):323–364.

Bonte, P., Turck, F. D., and Ongenae, F. (2022). Bridging the gap between expressivity and efficiency in stream reasoning: a structural caching approach for iot streams. *Knowledge and Information Systems*, 64(7):1781–1815.

Burgdorf, A., Paulus, A., Pomp, A., and Meisen, T. (2022). Docsemmap: Leveraging textual data documentations for mapping structured data sets into knowledge graphs. pages 209–216. IEEE.

Buron, M., Goasdoué, F., Manolescu, I., and Mugnier, M.-L. (2020). Obi-Wan: Ontology-Based RDF Integration of Heterogeneous Data. *Proceedings of the VLDB Endowment*, 13(12):2933–2936.

Calva, M. and Piedra, N. (2020). Health data representation through web semantic, a case study applied to electronic records medical in the utpl hospital. In *2020 XLVI Latin American Computing Conference (CLEI)*, pages 284–293, Loja, Ecuador. IEEE.

Capodieci, A., Mainetti, L., and Carrozzo, S. (2016). Semantic enterprise service bus for cultural heritage. In *2016 12th International Conference on Innovations in Information Technology (IIT)*, pages 1–8. IEEE.

Cheung, C. M., Goyal, P., Harris, G., Patri, O., Srivastava, A., Zhang, Y., Panangadan, A., Chelmis, C., McKee, R., Theron, M., Nemeth, T., and Prasanna, V. K. (2015). Rapid data integration and analysis for upstream oil and gas applications. In *SPE Annual Technical Conference and Exhibition*, pages 2573–2590, Houston, Texas, USA. SPE.

Cudré-Mauroux, P. (2020). Leveraging knowledge graphs for big data integration: the xi pipeline. *Semantic Web*, 11(1):13–17.

DAMA, I. (2017). *DAMA-DMBOK: Data Management Body of Knowledge (2nd Edition)*. Technics Publications, LLC, Denville, NJ, USA.

Dao, J., Ng, S. T., Yang, Y., Zhou, S., Xu, F. J., and Skitmore, M. (2021). Semantic framework for interdependent infrastructure resilience decision support. *Automation in Construction*, 130.

Dhayne, H., Kilany, R., Haque, R., and Taher, Y. (2018). Sedie: A semantic-driven engine for integration of healthcare data. In *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 617–622, Madrid, Spain. IEEE.

Ding, L., Xiao, G., Calvanese, D., and Meng, L. (2020). A framework uniting ontology-based geodata integration and geovisual analytics. *ISPRS International Journal of Geo-Information*, 9(8).

Dirgahayu, T., Hendrik, and Setiaji, H. (2020). Semantic web in disaster management: A systematic literature review. *IOP Conference Series: Materials Science and Engineering*, 803(1).

Divya, G. and Manish, T. I. (2020). Machine learning techniques and frameworks for heterogeneous data fusion in big data analytics. In *2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, pages 1568–1574. IEEE.

Dridi, A., Sassi, S., Chbeir, R., and Faiz, S. (2020). A flexible semantic integration framework for fully-integrated ehr based on fhir standard. In *ICAART 2020 - Proceedings of the 12th International Conference on Agents and Artificial Intelligence*, volume 2, pages 684–691, Valletta, Malta. SCITEPRESS - Science and Technology Publications.

Fathy, N., Gad, W., and Badr, N. (2019). A unified access to heterogeneous big data through ontology-based semantic integration. In *2019 Ninth International Conference on Intelligent Computing and Information Systems (ICICIS)*, pages 387–392. IEEE.

Fusco, G. and Aversano, L. (2020). An approach for semantic integration of heterogeneous data sources. *PeerJ Computer Science*, 6(3).

Gil, R. M., de Buenaga Rodríguez, M., Galisteo, F. A., Páez, D. G., and García-Cuesta, E. (2021). A domain-adaptable heterogeneous information integration platform: Tourism and biomedicine domains. *Information*, 12(11).

Gomes, L. and Santanchè, A. (2015). The web within: Leveraging web standards and graph analysis to enable application-level integration of institutional data. *Lecture Notes in Computer Science*, 8990:26–54.

Grasso, C. T., Joshi, A., and Siegel, E. (2015). Beyond ner: Towards semantics in clinical text. In *Proceedings of International Workshop on Biomedical Data Mining, Modeling, and Semantic Integration: A Promising Approach to Solving Unmet Medical Needs (BDM2I 2015)*, volume 1428, Bethlehem, PA, USA. CEUR Workshop Proceedings.

Gruber, T. R. (1993). A translation approach to portable ontology specifications. *Knowledge Acquisition*, 5(2):199–220.

Guarino, N. (1998). Formal ontology in information systems. pages 3–15.

Guedea-Noriega, H. H. and García-Sánchez, F. (2022). Integroly: Automatic knowledge graph population from social big data in the political marketing domain. *Applied Sciences*, 12(16).

Gupta, N. and Gupta, B. (2021). Machine learning approach of semantic mapping in polystore health information systems. *International Journal of Computer Information Systems and Industrial Management Applications*, 13:222–233.

Haghgoo, M., Mazary, A. N. A., and Monti, A. (2022). Siseg-auto semantic annotation service to integrate smart energy data. *Energies*, 15(4):1428.

Hao, X., Ji, Z., Li, X., Yin, L., Liu, L., Sun, M., Liu, Q., and Yang, R. (2021). Construction and application of a knowledge graph. *Remote Sensing*, 13(13).

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Jeong, H. and Jeong, H. (2015). Ontology-based integration and refinement of evaluation-committee data from heterogeneous data sources. *Indian Journal of Science and Technology*, 8(23):1–7.

Jie, F., Huang, Y., Bai, Q., and Wu, X. (2021). Hao unity: A graph-based system for unifying heterogeneous data. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, CIKM '21, page 4725–4729, New York, NY, USA. Association for Computing Machinery.

Kamm, S., Jazdi, N., and Weyrich, M. (2021). Knowledge discovery in heterogeneous and unstructured data of industry 4.0 systems: Challenges and approaches. *Procedia CIRP*, 104:975–980.

Katrandzhiev, K., Gocheva, K., and Bratanova-Doncheva, S. (2022). Whole system data integration for condition assessments of climate change impacts: An example in high-mountain ecosystems in rila (bulgaria). *Diversity*, 14(4).

Kessler, I., Perzylo, A., and Rickert, M. (2021). Ontology-based decision support system for the nitrogen fertilization of winter wheat. In E., O.-P. M. G., editor, *Metadata and Semantics Research*, volume 1355 of *Communications in Computer and Information Science*, pages 245–256, Madrid, Spain. Springer, Cham.

Khnaisser, C., Lavoie, L., Fraikin, B., Barton, A., Dussault, S., Burgun, A., and Ethier, J.-F. (2022). Using an ontology to derive a sharable and interoperable relational data model for heterogeneous healthcare data and various applications. *Methods of Information in Medicine*, 61(2):e73–e88.

Kim, J., Kong, J., Sohn, M., and Park, G. (2021). Layered ontology-based multi-sourced information integration for situation awareness. *The Journal of Supercomputing*, 77(9):9780–9809.

Kitchenham, B. A. and Charters, S. (2007). Guidelines for performing systematic literature reviews in software engineering. Technical Report EBSE 2007-001, Keele University and Durham University Joint Report.

Krataithong, P., Anutariya, C., and Buranarach, M. (2022). A taxi trajectory and social media data management platform for tourist behavior analysis. *Sustainability*, 14(8):4677.

Kumar, N. and Das, S. (2019). New distributed data fusion using pregel for large text dataset. *International Journal of Scientific and Technology Research*, 8(12):2090–2098.

Lafferty, J. D., McCallum, A., and Pereira, F. C. N. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. page 282–289.

Le Guillarme, N., Hedde, M., and Thuiller, W. (2021). Building a trophic knowledge graph to support soil food web reconstruction. In *S4BioDiv 2021: 3rd International Workshop on Semantics for Biodiversity*, volume 2969, Bolzano, Italy. CEUR Workshop Proceedings.

Lembo, D. and Scafoglieri, F. M. (2020). Ontology-based document spanning systems for information extraction. *International Journal of Semantic Computing*, 14(01):3–26.

Ma, B., Jiang, T., Zhou, X., Zhao, F., and Yang, Y. (2017). A novel data integration framework based on unified concept model. *IEEE Access*, 5:5713–5722.

Ma, C. and Molnár, B. (2020). Use of ontology learning in information system integration: A literature survey. In *Communications in Computer and Information Science*, volume 1178 CCIS, pages 342–353. Springer, Singapore.

Ma, C. and Molnár, B. (2022). Ontology learning from relational database: Opportunities for semantic information integration. *Vietnam Journal of Computer Science*, 09(01):31–57.

Maga-Nteve, C., Kontopoulos, E., Tsolakis, N., Katakis, I., Mathioudis, E., Mitzias, P., Avgerinakis, K., Meditskos, G., Karakostas, A., Vrochidis, S., and Kompatsiaris, I. (2022). A semantic technologies toolkit for bridging early diagnosis and treatment in brain diseases: Report from the ongoing eu-funded research project alameda. volume 1537 of *CCIS*, pages 349–354. Springer Science and Business Media Deutschland GmbH.

Mahmoud, A., Shams, M. Y., Elzeki, O. M., and Awad, N. A. (2021). Using semantic web technologies to improve the extract transform load model. *Computers, Materials & Continua*, 68(2):2711–2726.

Mami, M. N., Graux, D., Scerri, S., Jabeen, H., Auer, S., and Lehmann, J. (2019). Uniform access to multiform data lakes using semantic technologies. In *Proceedings of the 21st International Conference on Information Integration and Web-Based Applications & Services*, iiWAS2019, page 313–322. Association for Computing Machinery.

Masmoudi, M., Karray, M. H., Lamine, S. B. A. B., Zghal, H. B., and Archimede, B. (2019). Diserto: Semantics-based tool for automatic and virtual data integration. In *2019 IEEE/ACS 16th International Conference on Computer Systems and Applications (AICCSA)*, pages 1–8, Abu Dhabi, United Arab Emirates. IEEE.

Masseroli, M., Canakoglu, A., and Ceri, S. (2016). Integration and querying of genomic and proteomic semantic annotations for biomedical knowledge extraction. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 13(2):209–219.

Mountasser, I., Ouhbi, B., Hdioud, F., and Frikh, B. (2021). Semantic-based big data integration framework using scalable distributed ontology matching strategy. *Distributed and Parallel Databases*, 39(4):891–937.

Mrhar, K., Douimi, O., Abik, M., and Benabdellah, N. C. (2020). Towards a semantic integration of data from learning platforms. *IAES International Journal of Artificial Intelligence*, 9(3):535–544.

Nagpal, P., Chaudhary, D., and Singh, J. (2021). Knowing the unknown: Unshielding the mysteries of semantic web in health care domain. In *Proceedings of the Workshop on Advances in Computational Intelligence, its Concepts & Applications (ACI 2021)*, volume 2823, pages 37–44. CEUR Workshop Proceedings.

Nashipudimath, M. M., Shinde, S. K., and Jain, J. (2020). An efficient integration and indexing method based on feature patterns and semantic analysis for big data. *Array*, 7.

Nath, R. P. D., Hose, K., Pedersen, T. B., and Romero, O. (2017). Setl: A programmable semantic extract-transform-load framework for semantic data warehouses. *Information Systems*, 68:17–43.

Nathalie, A. (2009). Schema matching based on attribute values and background ontology. In *12th AGILE International conference on geographic information science*, volume 1, pages 1–9. Springer Berlin, Heidelberg.

Niang, C., Marinica, C., Élise Leboucher, Bouiller, L., Capderou, C., and Bouchou, B. (2016). Ontology-based data integration system for conservation-restoration data (obdis-cr). In *Proceedings of the 20th International Database Engineering & Applications Symposium*, IDEAS '16, page 218–223, New York, New York, USA. Association for Computing Machinery.

Nimmagadda, S. L., Reiners, T., and Wood, L. C. (2019). On modelling big data guided supply chains in knowledge-base geographic information systems. *Procedia Computer Science*, 159:1155–1164.

Noriega, H. H. G. and Sanchez, F. G. (2019). Semantic (big) data analysis: an extensive literature review. *IEEE Latin America Transactions*, 17(05):796–806.

Nundloll, V., Lamb, R., Hankin, B., and Blair, G. (2021). A semantic approach to enable data integration for the domain of flood risk management. *Environmental Challenges*, 3.

Oo, S. M., Haesendonck, G., Meester, B. D., and Dimou, A. (2022). Rmlstreamer-siso: An rdf stream generator from streaming heterogeneous data. In *International Semantic Web Conference*, pages 697–713.

Pereira, A., Lopes, R. P., and Oliveira, J. L. (2020). Scaleus-fd: A fair data tool for biomedical applications. *BioMed research international*.

Phengsuwan, J., Shah, T., Sun, R., James, P., Thakker, D., and Ranjan, R. (2022). An ontology-based system for discovering landslide-induced emergencies in electrical grid. *Transactions on Emerging Telecommunications Technologies*, 33(3).

Pomp, A., Paulus, A., Jeschke, S., and Meisen, T. (2017). Eskape: Platform for enabling semantics in the continuously evolving internet of things. In *2017 IEEE 11th International Conference on Semantic Computing (ICSC)*, pages 262–263, San Diego, CA, USA. IEEE.

Qundus, J. A., Schäfermeier, R., Karam, N., Peikert, S., and Paschke, A. (2021). Roc: An ontology for country responses towards covid-19. In *2nd International Conference on Digital Curation Technologies*, volume 2836, Berlin, Germany. CEUR Workshop Proceedings.

Radaoui, M., Ben Abdallah Ben Lamine, S., Zghal, H. B., Guegan, C. G., and Kabachi, N. (2019). Knowledge guided integration of structured and unstructured data in health decision process. In *Proceedings of the 28th International Conference on Information Systems Development: Information Systems Beyond 2020, ISD 2019*, Toulon, France: ISEN Yncréa Méditerranée.

Ramzy, N., Auer, S., Ehm, H., and Chamanara, J. (2022). Mare: Semantic supply chain disruption management and resilience evaluation framework. In *Proceedings of the 24th International Conference on Enterprise Information Systems*, volume 2, pages 484–493. SCITEPRESS - Science and Technology Publications.

Rani, P. S., Suresh, R. M., and Sethukarasi, R. (2019). Multi-level semantic annotation and unified data integration using semantic web ontology in big data processing. *Cluster Computing*, 22(5):10401–10413.

Rouces, J., de Melo, G., and Hose, K. (2018). Addressing structural and linguistic heterogeneity in the web1. *AI Communications*, 31(1):3–18.

Saber, A., Al-Zoghby, A. M., and Elmougy, S. (2018). Big-data aggregating, linking, integrating and representing using semantic web technologies. In Hassanien, A. E., Tolba, M. F., Elhoseny, M., and Mostafa, M., editors, *The International Conference on Advanced Machine Learning Technologies and Applications (AMLTA2018)*, volume 723, pages 331–342, Cairo, Egypt. Springer International Publishing.

Sandhya, H. and Roy, M. M. (2016). Data integration of heterogeneous data sources using qr decomposition. In S., T. S. M. B. S. D., editor, *International Symposium on Intelligent Systems Technologies and Applications*, volume 385 of *Advances in Intelligent Systems and Computing*, pages 333–344. Springer Verlag.

Santipantakis, G. M., Glenis, A., Patroumpas, K., Vlachou, A., Doulkeridis, C., Vouros, G. A., Pelekis, N., and Theodoridis, Y. (2020). Spartan: Semantic integration of big spatio-temporal data from streaming and archival sources. *Future Generation Computer Systems*, 110:540–555.

SCHIESSL, M. and BRÄSCHER, M. (2017). Ontology lexicalization: Relationship between content and meaning in the context of information retrieval. *Transinformação*, 29(1):57–72.

Sengloiluean, K. and Khuntong, R. (2020). Ontology-based semantic integration of heterogeneous data sources using ontology mapping approach. *Journal of Theoretical and Applied Information Technology*, 98(22):3489–3502.

Sernadela, P. and Oliveira, J. L. (2017). A semantic-based workflow for biomedical literature annotation. *Database*, 2017.

Shen, F., Liu, H., Sohn, S., Larson, D. W., and Lee, Y. (2016). Predicate oriented pattern analysis for biomedical knowledge discovery. *Intelligent Information Management*, 08(03):66–85.

Sima, A. C., de Farias, T. M., Zbinden, E., Anisimova, M., Gil, M., Stockinger, H., Stockinger, K., Robinson-Rechavi, M., and Dessimoz, C. (2019). Enabling semantic queries across federated bioinformatics databases. *Database : the journal of biological databases and curation*.

Stroganov, O., Fedarovich, A., Wong, E., Skovpen, Y., Pakhomova, E., Grishagin, I., Fedarovich, D., Khasanova, T., Merberg, D., Szalma, S., and Bryant, J. (2022). Mapping of uk biobank clinical codes: Challenges and possible solutions. *PLOS ONE*, 17(12).

Sugawara, E. and Nikaido, H. (2014). Properties of adeabc and adeijk efflux systems of acinetobacter baumannii compared with those of the acrab-tolc system of escherichia coli. *Antimicrobial agents and chemotherapy*, 58(12):7250–7.

Thirumahal, R., Sadasivam, G. S., and Shruti, P. (2022). Semantic integration of heterogeneous data sources using ontology-based domain knowledge modeling for early detection of covid-19. *SN Computer Science*, 3(6).

VandanaKolisetty, V. and Rajput, D. S. (2021). Integration and classification approach based on probabilistic semantic association for big data. *Complex & Intelligent Systems*.

Vidal, M.-E., Jozashoori, S., and Sakor, A. (2019). Semantic data integration techniques for transforming big biomedical data into actionable knowledge. In *2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS)*, pages 563–566. IEEE.

Vilches-Blázquez, L. M. and Saavedra, J. (2022). A graph-based representation of knowledge for managing land administration data from distributed agencies – a case study of colombia. *Geo-spatial Information Science*, 25(2):259–277.

Wache, H., Vögele, T., Visser, U., Stuckenschmidt, H., Schuster, G., Neumann, H., and Hübner, S. (2001). Ontology-based integration of information - a survey of existing approaches. In *Proceedings of the IJCAI-01 Workshop on Ontologies and Information Sharing*, pages 108–118. CEUR Workshop Proceedings.

Wang, X., Xu, J., Liu, M., Wei, Z., Bu, W., and Hong, T. (2017). An ontology-based approach for marine geochemical data interoperation. *IEEE Access*, 5:13364–13371.

Wu, J., Orlandi, F., O'Sullivan, D., and Dev, S. (2022). A workflow to convert live atmospheric sensor data into linked data. In *IGARSS 2022 - 2022 IEEE International Geoscience and Remote Sensing Symposium*, pages 4086–4089. IEEE.

Xiao, H. (2006). *Query Processing for Heterogeneous Data Integration Using Ontologies*. PhD thesis.

Xiao, W., Guoqi, L., and Bin, L. (2017). Research on big data integration based on karma modeling. In *IEEE International Conference on Software Engineering and Service Science*, pages 245–248, Beijing, China. IEEE.

Yadav, M., Singh, V., and Prachi (2021). Ontology based data integration and mapping for adverse drug reaction. In *2021 6th International Conference on Signal Processing, Computing and Control (ISPCC)*, pages 719–727, Solan, India. IEEE.

Yafooz, W. M. S., Abdin, S. Z., and Fahad, S. A. (2018). Managing textual data semantically in relational databases. In *2018 International Conference on Smart Computing and Electronic Enterprise (ICSCEE)*, pages 1–5, Shah Alam, Malaysia. IEEE.

Yun, H., He, Y., Lin, L., and Wang, X. (2019). Research on multi-source data integration based on ontology and karma modeling. *International Journal of Intelligent Information Technologies*, 15(2):69–87.

Zhang, H., Guo, Y., Li, Q., George, T. J., Shenkman, E., Modave, F., and Bian, J. (2018). An ontology-guided semantic data integration framework to support integrative data analysis of cancer survival. *BMC Medical Informatics and Decision Making*, 18(2).

Zhang, S., Tang, Y., Yan, J., Li, L., Li, T., Li, J., Xie, P., Gu, Y., Xu, J., Feng, Z., Zhang, W., Xia, J., Mayer, W., Zhang, H.-Y., He, G.-C., and He, K. (2021). A graph-based approach for integrating biological heterogeneous data based on connecting ontology. In *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 600–607, Houston, TX, USA. IEEE.

Zhao, Q., Liu, J., Sullivan, N., Chang, K. C., Spina, J., Blasch, E., and Chen, G. (2021). Anomaly detection of unstructured big data via semantic analysis and dynamic knowledge graph construction. In *Proc. SPIE 11756, Signal Processing, Sensor/Information Fusion, and Target Recognition XXX*. SPIE.

Zhao, W., Zhou, B., and Zhang, C. (2022). Heterogeneous social linked data integration and sharing for public transportation. *Journal of Advanced Transportation*, pages 1–14.

Zhou, Q. (2016). Research on heterogeneous data integration model of group enterprise based on cluster computing. *Cluster Computing*, 19(3):1275–1282.

Ziegler, P. and Dittrich, K. R. (2004). Three decades of data intecration — all problems solved? volume 156, pages 3–12. Springer US.

Özsu, M. T. and Valduriez, P. (2020). *Principles of Distributed Database Systems*. Springer International Publishing, 4 edition.