




Publishing and Consuming Semantic Views for Construction of Knowledge Graphs

Narciso Arruda¹^a, Amanda D. P. Venceslau¹^b, Matheus Mayron¹, V. M. P. Vidal¹
and V. M. Pequeno²^c

¹*Departamento de Computação, Federal University of Ceará, Fortaleza, Ceará, Brazil*

²*TechLab, Departamento de Ciências e Tecnologias, Universidade Autónoma de Lisboa Luís de Camões, Portugal*

Keywords: Knowledge Graph, Semantic Integration, Vocabulary, Data Quality, Semantic View.

Abstract: The main goal of semantic integration is to provide a virtual semantic view that is semantically connected to data so that applications can have integrated access to data sources through the virtual Knowledge Graph. A semantic view can be published on a semantic portal to make it reusable for building Knowledge Graphs for different applications. This paper takes the first step towards publishing a semantic view on a semantic portal. This paper has three main contributions. First, we introduce a vocabulary for specifying semantic views. Then, we introduce a vocabulary for specification and quality assessment of Knowledge Graph. Third, we describe an approach to automatize the construction of a high-quality Knowledge Graph reusing a semantic view.

1 INTRODUCTION

In recent years, with the increase in the amount of public data available, it has also increased the number of applications that demand large volumes of heterogeneous data in order to allow greater accuracy in data analysis. These types of applications require the creation of a homogeneous view of the data.


Recently, the term Knowledge Graph (KG) has been used in association with Semantic Web technologies, linked data, large-scale data analytics and cloud computing (Ehrlinger and Wöß, 2016). The use of knowledge graphs is on the rise as a way of building homogeneous knowledge bases as a graph structure, combine different heterogeneous databases.


In the literature, the virtual Knowledge Graph (VKG) approach has been discussed in a paradigm known as ontology-based data access (OBDA) (Xiao et al., 2018). The VKG approach proposes to use one consistent virtual graph, more flexible than a rigid table structure and embed domain knowledge (Xiao et al., 2019). In VKG, integration views are virtual, which enables simplified design and maintenance as these views can be tested and modified instantly. However, the amount of data encountered can


cause query performance bottlenecks and to remedy this problem, many organizations create copies of the data, also called specialized Knowledge Graph, materializing a portion of the data as required. In addition, specialized Knowledge Graph goes through a process of conflict resolution and the use of data quality metrics that allow queries over a consolidated database.

We call semantic integration the process that makes use of a conceptual representation of the data and its relationship to deal with heterogeneity. The main goal of semantic integration is to provide a virtual semantic view, which is semantically connected to data so that applications can have integrated access to data sources. To make the semantic view reusable, it should be published on a Semantic Portal, which is intended to consolidate and semantically integrate large numbers of heterogeneous data sources into a comprehensive dataspace. Chem2bio2rdf (Bleiholder and Naumann, 2010), SemanticDB (D Pierce et al., 2012), and SemanticSUS (da Cruz et al., 2019) are examples of semantic portals.

Once a semantic view (see, Section 2) has been specified and published, semantic integration has already been performed a priori, so it can be used to build a virtual Knowledge Graph or reused to build specialized Knowledge Graph using Mashup View (see, Section 3), which can also be published and accessed by external applications. The semantic view allows integration approaches like (Collarana et al.,

^a  <https://orcid.org/0000-0003-3873-8468>

^b  <https://orcid.org/0000-0003-4118-4224>

^c  <https://orcid.org/0000-0002-6424-0252>

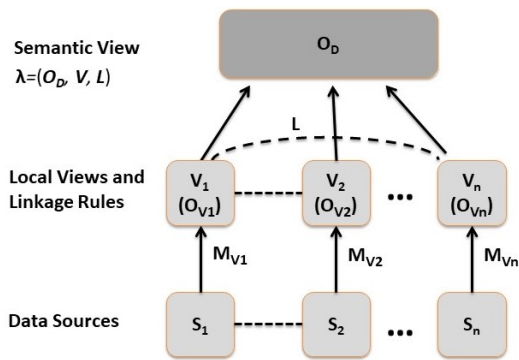


Figure 1: Three Level Framework for Semantic View Specification.

2017; Schultz et al., 2011), to be able to define the integration steps, providing the reuse and reduction of time-consuming for design other applications.

This paper takes the first step towards publishing a semantic view. The proposed approach combines ontologies and linked data to face the challenges in developing applications where there is a need to integrate heterogeneous data sources. The paper has three main contributions, which are described in sections 2, 3 and 4. Section 2 introduces a vocabulary for specifying a semantic view. A semantic view is specified with the help of local views and sameAs linkset views. Section 3 introduces a vocabulary for specification and quality assessment of Knowledge Graphs, providing relevant source information and quality metrics to benefit applications that aim to build high quality KG (Collarana et al., 2017). It also discusses how to reuse a semantic view specification for semi-automatic generation of a Knowledge Graph. Section 4 describes an approach for building high-quality Knowledge Graphs based on the quality metadata of a semantic view. Finally, Section 5 contains the conclusions.

2 SEMANTIC VIEW

In this section, we discuss a three-level ontology-based framework (Vidal et al., 2015), as summarized in Figure 1, to formally specify semantic view and a vocabulary to represent this specification and metadata.

2.1 Semantic View Specification

In the proposed framework, the semantic view resulting from semantic integration over data sources S_1, \dots, S_n is a triple $\lambda = (O_D, V, L)$, where:

- O_D represents the domain ontology (semantic view layer). O_D is responsible for establishing a vocabulary to be shared to describe the data sources;
- V represents a set of local view specifications V_1, \dots, V_n that describes the data sources S_1, \dots, S_n using the terms in O_D . A local view specification V_i is a tuple (O_{V_i}, M_{V_i}) , where:
 - O_{V_i} is the ontology of the local view. The vocabulary of O_{V_i} is a subset of the vocabulary of O_D whose terms occur in M_{V_i} .
 - M_{V_i} is a set of mappings that relate terms of vocabulary O_D with terms S_i ;
- L is a set of linkage rules that specify virtual sameAs links between resources in different local views. These links are used to relate resources that represent the same entity of the real world. We consider two types of sameAs links: *imported sameAs links*, which are exported by a Linked Data source, and *mashup sameAs links*, which are automatically created based on a sameAs linkset view specification (Casanova et al., 2014) specifically defined for the mashup application;

The process for generating the semantic view specification λ consists of 3 steps: (1) Modeling of the domain ontology; (2) Generation of the local views specifications; (3) Generation of the linkage rules.

2.2 Semantic View Vocabulary

The vocabulary for describing semantic view (VSV) is partitioned in three categories: General Metadata, View Specification Metadata and Quality Metadata. Given that a semantic view is a virtual dataset, the vocabulary for general metadata uses the terms in VoID vocabulary (Hartig and Zhao, 2010) for providing basic metadata about a dataset. VoID (prefix void:) is an RDF Schema vocabulary for expressing metadata about RDF datasets. It is intended as a bridge between the publishers and users of RDF data, with applications ranging from data discovery to cataloging and archiving of datasets. The main terms in VoID for general metadata are: *dcterms:description*, *dcterms:created*, *dcterms:license*, *dcterms:source*, *dcterms:Source* and *dcterms:vocabulary*.

The vocabulary for expressing metadata about a semantic view specification is defined as OWL ontology. Fig. 3 shows the main fragment of the VSV vocabulary (prefix vsv:). The main classes and properties are: *vsv:SemanticViewSpecification*, *owl:Ontology*, *vsv:LocalView*, *vsv:LinkageRule*, *vsv:QualityMetadata*, *vsv:hasQualityMetadata*, *vsv:hasLocalView*. The vocabulary was developed

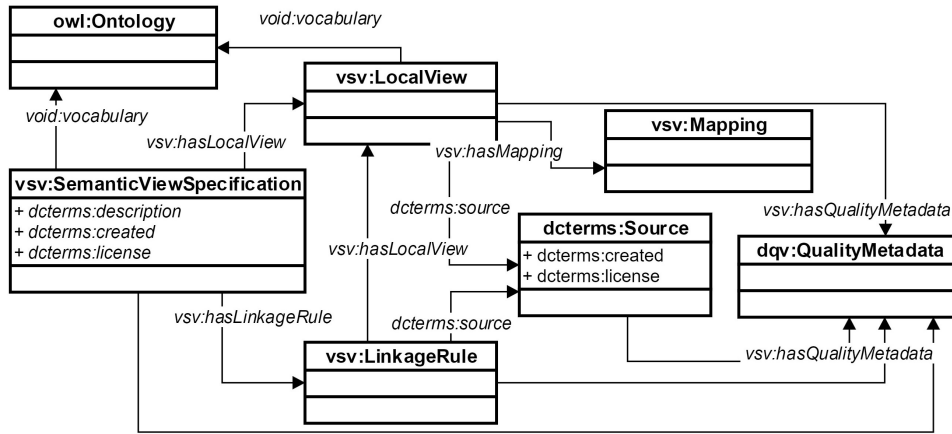


Figure 2: A Fragment of the Semantic View Specification Vocabulary.

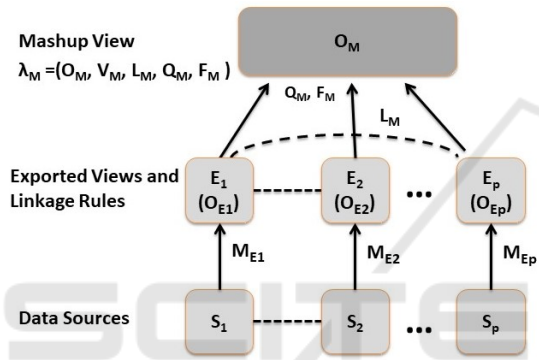


Figure 3: Three Level Framework for Mashup View Specification.

with understandability and usability in mind. For this reason, we apply a consistent scheme for property names, using "has" followed by a class name.

For expressing metadata about the quality of a semantic view, we use the terms in Data Quality Vocabulary (DQV) (prefix dqv:) (Debattista et al., 2016) discussed in Section 4. In our framework, the quality of semantic views is computed based on the quality of the local views and quality of the linkset Views. For more details see (Zaveri et al., 2013).

3 MASHUP VIEW

The creation of a mashup view is a complex task which involves four major challenges: (1) selection of the Linked Data sources that are relevant for the application; (2) extraction and translation of data from different, possibly heterogeneous data sources to a common vocabulary; (3) identification of links between resources in different Linked Data sources; (4) combination and fusion of multiple representations of the same real-world object into a single representation

and resolution of data inconsistencies to improve the quality of the data.

In this section, we present an ontology-based framework (Vidal et al., 2015) used in our approach to specifying a mashup view, and a vocabulary to represent this specification and metadata. The materialization of a mashup view is calling specialized Knowledge Graph, it is automatically processed based on its specification.

3.1 Mashup View Specification

We use a three level ontology-based framework Vidal et al. (2015), as summarized in Figure 3, to formally specify Knowledge Graph. The specification of a Knowledge Graph M is a quintuple, $\lambda = (O_M, V_M, L_M, F_M, Q_M)$, where:

- O_M is the mashup view ontology;
- V_M is a set of exported view specifications E_1, \dots, E_n that describes the data sources S_1, \dots, S_n using the terms in O_M . Each view E_i is a tuple (M_{Ei}, O_{Ei}) , where:
 - M_{Ei} is a set of rules that relate terms of vocabulary O_M with terms of vocabulary S_i ;
 - O_{Ei} is the ontology of the exported view E_i . The vocabulary of O_{Ei} is a subset of the vocabulary of O_M whose terms occur in M_{Ei} .
- L_M is a set of linked view specifications L_1, \dots, L_m between E_1, \dots, E_n . We consider two types of sameAs links: *imported sameAs links*, which are exported by a Linked Data source, and *mashup sameAs links*, which are automatically created based on a sameAs linkset view specification Casanova et al. (2014) specifically defined for the mashup application;

- Q_M is a set of quality requirements, which are requested by the user application;
- F_M is a set of fusion rules that specify how to resolve the problem of contradictory attribute values when combining multiple representations of the same real-world object into a single representation (canonical IRI).

The process for generating the mashup view specification, without reusing a semantic view specification, consists of 4 steps: (1) Modeling of the mashup view ontology; (2) Generation of the exported views specifications; (3) Generation of the exported sameAs linkset view specifications; (4) Definition of quality requirements and fusion rules. Note that, steps 2-4 requires semantic integration of the data source, which is not an easy task.

In case that a semantic view is previously specified, semantic integration is done a priori, and a mashup view specification $\lambda = (O_M, V_M, L_M, Q_M, F_M)$, can be automatically generated based on O_M , Q_M , and the semantic view specification $\lambda = (O_D, V, L)$. In this case, the vocabulary of O_M is a subset of the vocabulary of O_D , therefore O_M can be defined using a faceted search interface by selecting concepts and specifying filters.

3.2 Mashup View Vocabulary

The vocabulary for describing a mashup view (MV) is partitioned in three categories: General Metadata, Data Mashup View Specification Metadata and Quality Metadata. The vocabularies for general and quality metadata is the same one used by the semantic view and discussed in previous section. The vocabulary for expressing metadata about a mashup view specification is defined as an OWL ontology. Figure 4 shows a fragment of the MV vocabulary (prefix mv:). The main classes and properties are: *mv:MashupViewSpecification*, *owl:Ontology*, *mv:ExportedViewSpecification*, *mv:LinkageRule*, *dqv:QualityMetadata*, *mv:hasQualityMetadata*, *mv:hasExportedViewS*.

4 BUILDING HIGH-QUALITY KNOWLEDGE GRAPH

We start this section by presenting the main concepts of data quality and then discuss how to represent those concepts using the DQV vocabulary. We conclude this section by summarizing a data quality assessment methodology.

4.1 Data Quality Vocabulary

The standardized formulation of data quality in RDF/OWL facilitates transparency, verification, and sharing of linked data quality. Data on Web Best Practices (DWBP) point to the importance of publishing data quality information about data on the Web. For this purpose, the DWBP created a vocabulary to express data quality, called Data Quality Vocabulary (DQV) (Debattista et al., 2016).

Data quality is commonly conceived as a multi-dimensional construction with dimensions such as timeliness, completeness, consistency, interoperability, conciseness, representational conciseness and availability (Wang and Strong, 1996). The quality dimensions are composed of quality metrics, which measure the quality of the data along the dimensions (Bizer and Cyganiak, 2009). More specifically, quality metrics are heuristics designed to fit a specific assessment situation Wang (2005).

Figure 5 shows a fragment of the DQV vocabulary. The DQV vocabulary distinguishes between three layers of abstraction (metric, dimensions, and category), based on a survey presented in (Zaveri et al., 2016). Quality metrics (*dqv:Metric*) are grouped into quality dimension (*dqv:Dimension*), by property *dqv:inDimension*. Quality dimensions are grouped into quality category (*dqv:Category*), by property *dqv:inCategory*. *dqv:QualityMeasurement* represents a quality metric measure of a given resource (*rdfs:Resource*), a resource can be a set of data, a set of links, a graph or a set of triples in which quality measurement is performed.

Many quality metrics have been proposed to assess the quality of Linked Data sources because of the importance and use of this data (Zaveri et al., 2016). For example, in Table 1 the metrics M8 and M9 assess the quality of the interoperability dimension and the metrics M5, M6, and M7 assess the quality of the accessibility category.

4.2 Materialization and Quality Assessment of Knowledge Graph

In this section, we propose a method for generation and quality assessment of knowledge graph. To be useful, a knowledge graph must have good quality. Quality assessment of knowledge graph is not a simple process, as it involves other factors such as quality of data sources, quality of mappings, and quality of sameAs links. Normally, KG quality is calculated along at least three dimensions (Dong and Naumann, 2009): completeness, conciseness, and consistency. Consistency expresses how much the data are in the

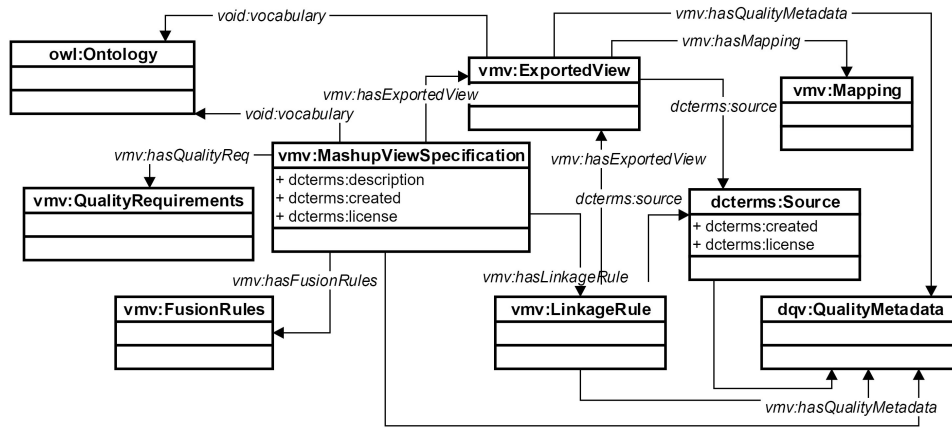


Figure 4: A Fragment of the Mashup View Specification Vocabulary.

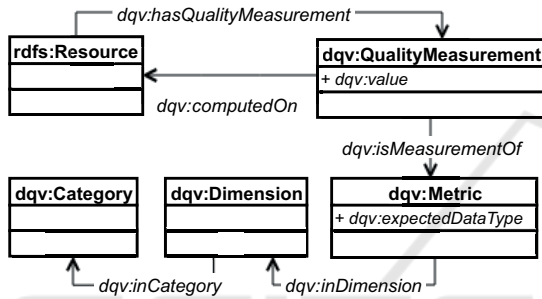


Figure 5: A Fragment of the **DQV** Vocabulary.

real world, while completeness and conciseness are, in a way, analogous to recall and precision in information retrieval (Knap et al., 2012).

As shown in Figure 6, materialization is performed incrementally, and at each step, the quality of the triples and datasets (materialized views) generated is also computed. The process of computing data quality is called quality assessment, in which process quality metadata is computed to measure data quality. Thus, errors can be detected by directing modifications that increase the quality of the data.

Table 2 shows quality metrics of exported view, linkset view and mashup view used to quality assessment of consistency dimension. The consistency of the mashup view is computed by the quality metrics MV_M1, MV_M2, and by the consistency of the instances of mashup view, that is computed by the consistency of its triples. Which in turn is computed by the metrics MV_M3 and MV_M4, and by the consistency of the exported views and linkset views. The consistency of the exported views and linkset views are computed by the metrics LV_M1, LV_M2 and metrics EV_M1, EV_M1, respectively. They depend on the consistency of the data source, the consistency of the linkset view also depends on the consistency of the exported views.

The generation of KG is processed automatically using as input the mashup view specification, data sources and quality metadata from data sources. As shown in Figure 6, the process of building the specialized knowledge graph consists of 3 steps, described in following.

Step 1. Materialization and Quality Assessment of Exported Views

In this step, each view in is materialized using the mappings. In this step, the quality of the exported view is also computed based on the quality metadata of the data sources, mapping rules, and the exported view materialization.

Step 2. Materialization and Quality Assessment of Linkset Views

This step identifies and materializes sameAs links. Each view is materialized using the linkage rule. Due to the importance of sameAs links, various approaches have been proposed to compute link quality, for example, based on functional properties (Papaleo et al., 2014) and using network measurements (Guéret et al., 2012)

Step 3. Data Fusion and Quality Assessment of Knowledge Graph

In this step, the fusion of multiple representations representing the same real-world entity into a single representation is performed. Fusion rules in F define how to solve the problem of conflicts that can occur in fusion objects. Resolving data inconsistency improves the quality of knowledge graph.

In the proposed framework, the specification of quality requirements, with the help of the user, should help in choosing which function to use to resolve a particular type of conflict. For example, if the user opts for a more complete mashup view, conflicts between values are not resolved.

As shown in Figure 6, during the data fusion process, the quality assessment of the generated triples

Table 1: Examples of Metrics, Dimensions and Categories.

Category	Dimension	Metric
Intrinsic	Consistency	M1 (Usage of incorrect domain or range data type)
		M2 (Misuse owl:DatatypeProperty or owl:ObjectProperty)
		M3 (Entities as members of disjoint classes)
Accessibility	Conciseness	M4 (Provides a measure of the redundancy of the dataset)
	Availability	M5 (deserferentiability of the URI)
		M6 (SPARQL endpoint availability)
Representational	Interoperability	M7 (RDF dump availability)
		M8 (existing terms reuse)
	Concision	M9 (existing vocabulary reuse)
		M10 (short URIs)

Table 2: Quality Metrics of the Consistency Dimension for Knowledge Graph.

Factor	Metric	Description
Mashup View	MV_M1	conformance of the source ontology and mashup ontology (schema consistency) (Wang (2012))
	MV_M2	mappings conforms to the semantics of information represented (mapping consistency) (Wang (2012))
	MV_M3	difference between value v and other (conflicting) values (Knap et al. (2012))
	MV_M4	confirmation values (Knap et al. (2012))
Exported View	EV_M1	The degree to which exported ontology is free of (logical/formal) contradictions (Zaveri et al. (2016))
	EV_M2	Proportion of mappings in the exported view error-free (Zaveri et al. (2016))
Linkset View	LV_M1	measures the similarity of instances linked to sameAs based on functional properties (Papaleo et al. (2014))
	LV_M2	measures the similarity of instances linked to sameAs based on linkage in linkset view

is performed. The quality of export views and links are important in determining the quality of the triple, also taking into account the equality and similarity of conflicting values (Knap et al., 2012).

5 CONCLUSIONS AND FUTURE WORK

This paper introduces a framework to publishing virtual Knowledge Graph on a semantic portal. First, we introduce a vocabulary for specifying semantic views. Then, we introduce a vocabulary for specification and quality assessment of Data Mashup view. Third, we describe an approach to automatize the construction of a high-quality Knowledge Graph reusing a semantic view specification. The proposed vocabularies as provides metadata for describing quality information about the semantic view and the mashup view. The quality information provided by the proposed vocabulary, enables the data quality assessment.

As a case study, we built SemanticSUS¹, a semantic portal which is intended to offer a semantic view that semantically integrates data sources from the unified health system of Brazil (SUS). In its current state, SemanticSUS semantically integrates three SUS data sources which are available on the GISSA platform (Freitas et al., 2017). The portal semantic View was used to generate the specification of the knowledge graph NDR (Neonatal Death Risk). The RMN mashup integrates information about children who lived less than 28 days (neonatal period), and it was used to develop a predictive model to establish the risk of neonatal death.

As a suggestion for future work, new metrics can be incorporated into the quality vocabulary. In (Aruda et al., 2019) we propose a Fuzzy evaluation approach to allow the creation of semantic rules (close to spoken language) to relate and evaluate the quality of the linked data. So using this fuzzy approach contribute to a more comprehensive quality assessment.

¹<https://semanticsus.github.io/semanticSUS/index.html>

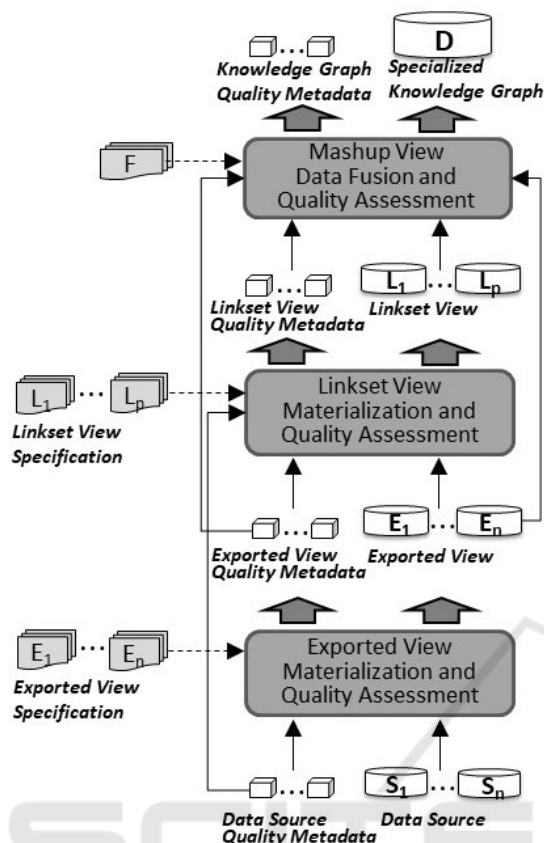


Figure 6: Knowledge Graph Materialization.

We are also developing a tool to facilitate the process of building high-quality Data Mashup views and incremental maintenance by reusing a Semantic view specification (Arruda, 2019). In our approach, the Data Mashup view is generated in three steps: first, the user specifies the mashup view ontology and the quality requirements of the mashup application. Then, the specification of the mashup view is automatically generated by reusing the mapping and linkage rules defined by the semantic view specification. Finally, the materialization and quality assessment of the Data Mashup view is automatically accomplished using the strategy described in Section 4.

REFERENCES

- Arruda, N. (2019). Framework for construction and incremental maintenance of high-quality linked data mashup. In *International Conference on Conceptual Modeling*, pages 213–221. Springer.
- Arruda, N., Alcântara, J., Vidal, V., Brayner, A., Casanova, M., Pequeno, V., and Franco, W. (2019). A fuzzy approach for data quality assessment of linked datasets. In *Proc. of the 21st Int. Conf. on Enterprise Information Systems*.
- Bizer, C. and Cyganiak, R. (2009). Quality-driven information filtering using the wiqa policy framework. *Journal of Web Semantics*, 7(1):1–10.
- Bleiholder, J. and Naumann, F. (2010). *Data fusion and conflict resolution in integrated information systems*. PhD thesis, University of Potsdam.
- Casanova, M. A., Vidal, V. M., Lopes, G. R., Leme, L. A. P. P., and Ruback, L. (2014). On materialized sameas linksets. In *International Conference on Database and Expert Systems Applications*, pages 377–384. Springer.
- Collarana, D., Galkin, M., Traverso-Ribón, I., Lange, C., Vidal, M.-E., and Auer, S. (2017). Semantic data integration for knowledge graph construction at query time. In *2017 IEEE 11th International Conference on Semantic Computing (ICSC)*, pages 109–116. IEEE.
- D Pierce, C., Booth, D., Ogbuji, C., Deaton, C., Blackstone, E., and Lenat, D. (2012). Semanticdb: a semantic web infrastructure for clinical research and quality reporting. *Current Bioinformatics*, 7(3):267–277.
- da Cruz, M. M. L., Avila, C. V. S., Vidal, V. M. P., and Junior, N. M. A. (2019). Semanticus: Um portal semântico baseado em ontologias e dados interligados para acesso, integração e visualização de dados do sus. In *Anais Estendidos do XIX Simpósio Brasileiro de Computação Aplicada à Saúde*, pages 13–18. SBC.
- Debattista, J., Auer, S., and Lange, C. (2016). Luzzu—a methodology and framework for linked data quality assessment. *Journal of Data and Information Quality (JDIQ)*.
- Dong, X. L. and Naumann, F. (2009). Data fusion: resolving data conflicts for integration. *Proceedings of the VLDB Endowment*, 2(2):1654–1655.
- Ehrlinger, L. and Wöß, W. (2016). Towards a definition of knowledge graphs. *SEMANTiCS (Posters, Demos, SuCESS)*, 48.
- Freitas, R., Rocha, C., Braga, O., Lopes, G., Monteiro, O., and Oliveira, M. (2017). Using linked data in the data integration for maternal and infant death risk of the sus in the gissa project. In *Proceedings of the 23rd Brazilian Symposium on Multimedia and the Web*, pages 193–196. ACM.
- Guéret, C., Groth, P., Stadler, C., and Lehmann, J. (2012). Assessing linked data mappings using network measures. In *Extended semantic web conf*. Springer.
- Hartig, O. and Zhao, J. (2010). Publishing and consuming provenance metadata on the web of linked data. In *International Provenance and Annotation Workshop*, pages 78–90. Springer.
- Knap, T., Michelfeit, J., Daniel, J., Jerman, P., Rychnovský, D., Soukup, T., and Nečaský, M. (2012). Od-cleanstore: a framework for managing and providing integrated linked data on the web. In *Int. Conf. Web Information Systems Engineering*, pages 815–816.
- Papaleo, L., Pernelle, N., and Saïs, F. (2014). On evaluating the quality of rdf identity links in the lod. In *In the proceedings of IC'2014 Workshop" From Open Sources to Web of Data"(SoWeDo 2014)*.
- Schultz, A., Matteini, A., Isele, R., Bizer, C., and Becker, C. (2011). LDIF - linked data integration framework. In

- Proc. Second Int. Conf. on Consuming Linked Data, COLD'11, Aachen, Germany. CEUR-WS.org.*
- Vidal, V. M., Casanova, M. A., Arruda, N., Roberval, M., Leme, L. P., Lopes, G. R., and Renso, C. (2015). Specification and incremental maintenance of linked data mashup views. In *International Conference on Advanced Information Systems Engineering*, pages 214–229. Springer.
- Wang, J. (2012). *A framework and architecture for quality assessment in data integration*. PhD thesis, University of London.
- Wang, R. Y. (2005). *Information Quality (Advances in Management Information Systems)*. M. E. Sharpe, Inc., Armonk, NY, USA.
- Wang, R. Y. and Strong, D. M. (1996). Beyond accuracy: What data quality means to data consumers. *J. Manage. Inf. Syst.*, 12(4):5–33.
- Xiao, G., Calvanese, D., Kontchakov, R., Lembo, D., Poggi, A., Rosati, R., and Zakharyashev, M. (2018). Ontology-based data access: A survey. *IJCAI*.
- Xiao, G., Ding, L., Cogrel, B., and Calvanese, D. (2019). Virtual knowledge graphs: An overview of systems and use cases. *Data Intelligence*, 1(3):201–223.
- Zaveri, A., Kontokostas, D., Sherif, M. A., Böhmann, L., Morsey, M., Auer, S., and Lehmann, J. (2013). User-driven quality evaluation of dbpedia. In *Proceedings of the 9th International Conference on Semantic Systems*, pages 97–104. ACM.
- Zaveri, A., Rula, A., Maurino, A., Pietrobon, R., Lehmann, J., and Auer, S. (2016). Quality assessment for linked data: A survey. *Semantic Web*, 7(1):63–93.

