International Journal of Information Technology and Language Studies (IJITLS)

A Hybrid Framework for Applying Semantic Integration Technologies to Improve Data Quality

Mahmoud Esmat Hamdy and **Khaled Shaalan**
mahmoudesmat@hotmail.com; khaled.shaalan@buid.ac.ae

The British University in Dubai, Dubai, UAE

Abstract. This study aims to develop a new hybrid framework of semantic integration for enterprise information system in order to improve data quality to resolve the problem from scattered data sources and rapid expansions of data. The proposed framework is based on a solid background that is inspired by previous studies. Significant and seminal research articles are reviewed based on selection criteria. A critical review is conducted in order to determine a set of qualified semantic technologies that can be used to construct a hybrid semantic integration framework. The proposed framework consists of six layers and one component as follows: source layer, translation layer, XML layer, RDF layer, inference layer, application layer, and ontology component. The proposed framework faces two challenges and one conflict; these were fixed while composing the framework. The proposed framework was examined to improve data quality for four dimensions of data quality dimensions.

Keywords: Semantic technology, data quality, hybrid framework.

1. Introduction

1.1 Semantic Integration

Semantic integration is a combination of data integration technology and semantics. The key for data integration is its ability to manipulate the data transparently across multiple sources (Cruz & Xiao, 2005). Regarding semantics, it can be defined as "the branch of linguistics and logic concerned with meaning" (Brouwer, 2016). Hence, when semantics and data integration are combined, this results in a process which uses a representation of data in a conceptual manner alongside the conceptual representation of the bonding or relationships which results in eliminating possible heterogeneities (Cruz & Xiao, 2005).

1.2 Heterogeneity Problems

The need for semantic integration is a result of data heterogeneity. Figure 1 depicts the types of heterogeneity problems (Brouwer, 2016).

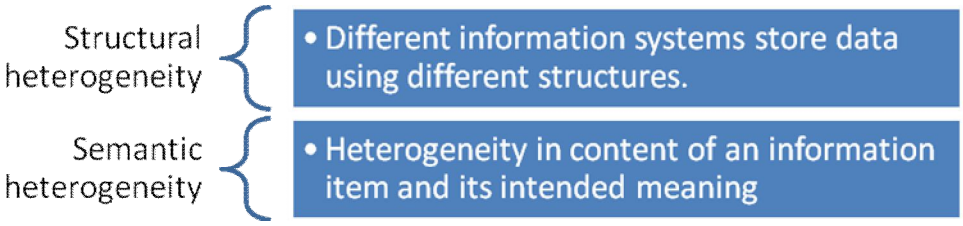

Figure 1. Types of Heterogeneity problems.

Goh (1997) summarizes the reasons for semantic heterogeneity. The reasons are listed below:

- Naming Conflicts: consists of synonyms and homonyms among attribute values.

- Scaling and units conflicts: adoption of different unit's measures or scales in reporting.
- Confounding conflicts: arises from confounding of concepts which are in fact distinct.
- Ontology is responsible for resolving data heterogeneity by achieving data interoperability. Ontology is defined as "specification of a conceptualization" (Gruber, 1993).

1.3 Data Quality

To achieve data quality (DQ), data quality is cascaded to data quality dimension. The data quality dimension is defined as "attributes that represent a single aspect or construct of data quality" (Wand & Wang 1996). Data quality dimensions are categorized as completeness, timeliness, accuracy, consistency, validity and uniqueness," (Askham et al., 2013). The data quality dimensions are used as benchmarks to assess data quality and measure the improvement.

1.4 Motivation

Different data type (structured, semi-structured, and unstructured) or data sources are rapidly increasing. For instance, data generated from new technologies like data from RFID technologies for inventory management, or data generated from autonomous systems, so the need to establish communication between them became major requirement. This motivates us to search about how to build bridges (connections) between various data sources. The essential keyword for that connection is the integration technologies, especially semantic integration. The reason for opting semantic integration is that the semantic integration is intelligent methodology more than other integration methodologies. There are many semantic integration technologies applied in various domains, and the consequences of applying semantic integration on data improve the level of data quality for data consumption.

We separate semantic integration technologies to main four areas: approach, framework, technique, and challenge. Many of current semantic integration technologies work to solve integration challenges for one type of dataset or data source. Some of these technologies are measured by the data quality improvement that is caused by applying these technologies. So, can we develop a new framework using current semantic integration technologies to improve data quality for various data sources or datasets? We split the initial question into sub-questions as follows:

What are the semantic technologies applied currently?

What are the semantic technologies approaches applied currently?

What are the semantic technologies frameworks applied currently?

What are the semantic technologies techniques applied currently?

What are the semantic technologies challenges applied currently?

How to perform data quality improvement and how to measure the data quality?

What is the proper measurement to measure the data quality?

What is the impact of applying semantic integration technologies on data quality improvement?

How to determine new integrated semantic technologies methodology that is suitable for improving data quality for enterprise information systems?

2 Literature Review

We start the literature review by sorting out a method to find the proper research articles by categorizing the requested research articles into the following categories:

- Research articles related to a specific domain: to provide us the relationship between theoretical concepts, methodologies, and practical application on business domain.
- Survey research articles: to provide a condensed overview of semantic integration technologies and data quality from previous work.
- Semantic integration technologies: to review research articles related to semantic integration approaches, frameworks, techniques, and challenges.
- Data quality: to review research articles related to data quality assessment.

Then we start surveying for all possible articles through search tools provided by the university's library and by google scholar. We used main key words to find possible articles like "Semantic integration", "Data quality", "Ontology", "Data quality dimensions", "Improve data Quality", "Semantic

integration challenges", "Schema matching" and many other key words. Some parts of the literature review were prepared according to the previous review studies conducted by (Al Emran & Shaalan, 2014; Al-Emran, 2015; Salloum et al., 2016; Salloum et al., 2017d; Salloum et al., 2018a; Salloum et al., 2018b) and some other studies carried out by (Al-Emran & Malik, 2016; Al-Emran et al., 2015; Al-Qaysi & Al-Emran, 2017; Mhamdi et al., 2018; Salloum et al., 2017a; Salloum et al., 2017b; Salloum et al., 2017c; Salloum et al., 2017e; Zaza & Al-Emran, 2015).

We screened one hundred thirty-six research articles and two books generated based on keywords criteria. Of these, we selected thirty-five research articles and two books for the literature review based on the following criteria:

- The research articles related to at least one of research questions.
- The research articles have been published in high ranked journals and got a good number of citations.
- The article contains proper description of the application.

We classify the selected research articles based on the following steps:

- Classify selected research articles based on main categories.
- Mapping each research article to one of the following sub-categories: the corresponded question to the article, conceptual or practical, the related domain and sort them chronologically in each category.

2.1 Semantic integration for specific domain

We selected semantic integration for Enterprise Information System as this field is emerging field and new data technology added to it frequently. Prabhakaran et al. (2006) highlighted the main critical success factor for Enterprise information system that is Master Data Management, Metadata, Enterprise-Wide Data Warehouse and Service Oriented Architecture. Prabhakaran et al. (2006), proposed to use ontology by using OWL language to create an additional layer for semantic integration. Based on the addition of semantic integration, Prabhakaran et al. (2006) developed, layout consisting of five layers.

Guido & Paiano (2010) proposed architecture by applying central integration point concept using Global as View approach. Guido and Paiano (2010) constructed global schema by the shared ontology. Finally, Guido and Paiano (2010) proposed a new methodology for ontology matching consisting of the following phases "linguistic normalization; semantic disambiguation; semantic similarity and tree matching".

2.2 Related Surveys

We selected survey research articles from different points of view; from the area that has rich amount of research articles, technology, challenges, and data quality. We selected research articles related to health information system for an area with rich amount of research articles. In health information system, a Survey article proposed by Liaw et al. (2013), surveyed the research papers introduced between 2001 till 2011 for improving data quality by semantic integration for chronic disease management. Out of two hundred forty-five papers screened they selected thirty-three papers, Liaw et al. (2013) developed new conceptual framework to review selected papers. Finding of Liaw et al. (2013) work can be concluded as four main points, they are:

- General and methodological
- Definitions of DQ and its operationalization and measurement in various studies
- Documented uses of ontologies for DQ, documented uses of ontology in CDM
- Documented uses of ontology in CDM

The first point concludes the reasons for exclusion of total papers poll and statistics about research questions (Liaw et al., 2013). The second point concludes the data quality dimensions related to the business domain. Liaw et al. (2013) found seven main data quality dimensions including completeness, consistency, correctness, accuracy (sub of correctness), reliability (sub of correctness), timeliness, relevance, usability, and security. The third point from the list, categorize semantic interoperability by ontology into three categories, the first category is based on the description of data quality. The second category is based on an assessment of data quality. The last category is for data quality management. The fourth point in the above list categorizes the uses of ontologies into three categories which include a description, management, and assessment.

From new technologies perspective, we select linked open data technology as that field is a new field and improve data quality is important for it. The selected article proposed by (Zaveri et al., 2016) surveyed twenty-one articles out of sixty-eight articles related to quality assessment for linked open data. (Zaveri et al., 2016) followed a methodology which involves five steps to select the proper papers, the methodology steps are as follows:

- Scan article titles based on inclusion/exclusion criteria
- Import to Mendeley and remove duplicates
- Review Abstracts and include/exclude articles
- Retrieve and analyze papers from references
- Compare Potentially shortlisted articles among reviewers

Zaveri et al. (2016) concluded their findings by formalizing the terminologies of data quality, finding out and defining the data quality dimensions and concluding the comparison between approaches from selected papers. For the first finding, Zaveri et al. (2016) formalized the meaning of the following terminologies:

- Data quality
- Data quality problems
- Data quality dimensions and metrics
- Data quality assessment methods.

The second finding, Zaveri et al. (2016) concluded the quality dimensions as follows:

- Accessibility dimensions
- Intrinsic dimensions
- Trust dimensions
- Dataset dynamicity dimensions
- Contextual dimensions
- Representational dimensions

The third finding shows the comparison between eight approaches from the selected articles; the approaches comparison was based on eight comparison criteria extracted from data quality dimensions (Zaveri et al. 2016). From challenges perspective, we focused on challenges related to ontology matching as that challenge one of the major challenge for semantic integration; Shvaiko and Euzenat (2013) conducted a review for ontology matching challenges and ontology matching challenge solutions.

Shvaiko and Euzenat (2013) concluded the ontology matching challenges into eight challenges:

- Large-scale evaluation
- The efficiency of ontology matching
- Matching with background knowledge
- Matcher selection and self-configuration
- User involvement
- Explanations of ontology matching
- Collaborative and social ontology matching
- Alignment infrastructure

Shvaiko and Euzenat (2013) compared between ontology matching solutions, table 1 illustrate the comparison between solutions.

System	Input	Output	GUI	Operation	Terminological	Structural	Extensional	Semantic
SAMBO §4.1	OWL	1:1 alignments	Yes	Ontology merging	n-gram, edit distance, UMLS, WordNet	Iterative structural similarity based on *is-a, part-of* hierarchies	Naive Bayes over documents	-
Falcon §4.2	RDFS, OWL	1:1 alignments	-	-	I-SUB, Virtual documents	Structural proximities, clustering, GMO	Object similarity	-
DSsim §4.3	OWL, SKOS	1:1 alignments	AQUA Q/A [31]	Question answering	Tokenization, Monger-Elkan, Jaccard, WordNet	Graph similarity based on leaves	-	Rule-based fuzzy inference
RiMOM §4.4	OWL	1:1 alignments	-	-	Edit distance, vector distance, WordNet	Similarity propagation	Vector distance	-
ASMOV §4.5	OWL	n:m alignments	-	-	Tokenization, string equality, Levenstein distance, WordNet, UMLS	Iterative fix point computation, hierarchical, restriction similarities	Object similarity	Rule-based inference
Anchor-Flood §4.6	RDFS, OWL	1:1 alignments	-	-	Tokenization, string equality, Winkler-based sim., WordNet	Internal, external similarities; iterative anchor-based similarity propagation	-	-
AgreementMaker §4.7	XML, RDFS, OWL, N3	n:m alignments	Yes	-	TF-IDF, edit distance, substrings, WordNet	Descendant, sibling similarities	-	-

Table 1. Comparison between ontology matching solutions (Shvaiko & Euzenat, 2013)

From data quality dimension perspective, a survey article introduced by (Weiskopf & Weng 2013) surveyed ninety-five papers, out of out of two hundred thirty papers screened regarding data quality assessment. Weiskopf and Weng (2013) found that most of papers reviewed, 73% covered only structured data or a combination of unstructured and structured data being 22%. Regarding data quality dimension terms, Weiskopf and Weng (2013) concluded that the papers included 27 unique terms which describes the dimensions of data quality Weiskopf and Weng (2013).

2.3 Semantic Integration Technologies

Integrate heterogeneity datasets, or data sources are a major fundamental problem for semantic integration because of complexity to identify that the data contains semantic information. The semantic information identified from the data refers to the real-world concept and can be integrated. There are many technologies used for semantic integration along to fix the challenges face applying it. This section will discuss approaches, frameworks, techniques and related challenges for semantic integration.

2.3.1 Approaches

There are two main methodologies for modeling the semantic integration schema related to semantic integration approaches. The methodologies are Global as view (GAV) and (LAV) (Lenzerini, 2002). In global as view "every element of the global schema is associated with a view, i.e., a query, over the sources, so that its meaning is specified in terms of the data residing at the sources" (Calì et al., 2005). In local as view "the sources is specified in terms of the elements of the global schema: more exactly, the mapping between the sources and the global schema is provided in terms of a set of views over the global schema, one for each source element" (Calì et al., 2005).

There are two main approaches for semantic integration. The approaches are central data management integration system and peer to peer data management integration system (Cruz & Xiao, 2005). "A central data integration system usually has a global schema, which provides the user with a uniform interface to access information stored in the data sources" (Cruz & Xiao, 2005). Peer to peer data management integration system can be defined as integration without centralized point marking as global point for integration as any peer can communicate (Cruz & Xiao, 2005). Peer to Peer integration was traditionally designed by first-order logic technique. The first order technique was criticized and proof poor

integration (Calvanese et al., 2003). Calvanese et al. (2003) used another technique named epistemic logic to achieve rich and proper integration. Calvanese et al. (2003) aim the modular structure to connect the different peers by the proper semantics. The proposed approach by Calvanese et al. (2003), consists of three main components. The components are framework, semantic and query answering (Calvanese et al., 2003). The designed framework shared set of constants to all peers, then the set of constants were related to first-order logic query (Calvanese et al., 2003).

A new semantic was designed as an enhancement of first-order logic by using epistemic logic which extends the apriori of the topology to additional peers (Calvanese et al., 2003). Calvanese et al. (2003) add more restrictions to query answering as the framework affect the answering. The language used for the query answering is the union of conjunctive queries, the query answering supports polynomial-time data regardless of the size of data. Another approach for peer to peer sematic integration was introduced by Cruz et al. (2004), to establish integration between two types of data sources (XML, RDF). Both the sources are totally different as XML is based on document structure while RDF is based on concepts and relations. To resolve the heterogeneous between both the data sources, Cruz et al. (2004) proposed new framework based on peer to peer integration approach as infrastructure for semantic central data integration approach. Global as view methodology was used. Alongside the approach considered one peer as super peer if it has the ontology and other peers. The process of mapping has two main sub-processes, map local RDF schema to global ontology and map local XML schema to global ontology (Cruz et al. 2004). The query processing has been designed by the following languages Xquery and RDQL, the processing run into two modes, that is integration mode and hybrid p2p mode (Cruz et al., 2004). Cruz et al. (2004) used hybrid approach of semantic integration to integrate different source formats, Dimartino et al. (2015) proposed pure peer to peer semantic integration to integrate open linked data from the same source format (RDF). Dimartino et al. (2015) propose a new semantic integration framework for triple tired integration architecture. The architecture of Dimartino et al. (2015) is more than the traditional method of double integration architecture, for each peer there is a corresponded schema related to the peer. Each schema consists of a group of URI matched with the model data after this the mapping has been created between groups of URI. Dimartino et al. (2015) designed RDF peer system containing a group of mapping and a group of peers which define the relationship between peers. Then RDF peer system semantic take place in a stored database. The query answering has been designed based on first-order logic to achieve the relational triples RDF. However, answering faced the problem of data exchange, which required to evaluate the query by universal solutions. Universal solution is a technique used to track dependencies in a database.

Due to universal solution being a clear miss, an additional step was amended to query answering which is query rewriting. As a latest a solution for problems related to peer to peer semantic data integration approach. Caroprese and Zumpano (2017) proposes solution to use deductive database in semantic integration of peer to peer integration approach. Database of peer to peer semantic integration considered deductive database (a combination between rational database and logic programing) (Caroprese & Zumpano, 2017). The peer to peer data integration rely upon mapping rules, there is a problem of interpretations of mapping rules, the proposed solution proposes a new semantic to enhance interpretations of mapping rules by using integrity constraints (Caroprese & Zumpano, 2017). With this the mapping rules retrieve maximum group of facts from the consistent peer, the technique was named as Preferred Weak Model (Caroprese & Zumpano, 2017). In other words, the model proposed by Caroprese and Zumpano (2017), solve the problem of each peer prefer the corresponded local schema to him, by categorizing peers as sound or unsound peer. Peer sound is where “peer can declare either its preference to its own knowledge” and unsound peer “give no preference to its own knowledge with respect to imported knowledge” (Caroprese & Zumpano, 2017).

2.3.2 Framework

The following section reviews various semantic integration frameworks. “Framework is a reusable design and building blocks for a software system and/or subsystem” (Shan & Hua, 2006). The standard semantic integration framework (TSIF) consists of three main components. The first main component of TSIF is the Global Schema. “Global schema provides a reconciled, integrated and virtual view of the underlying sources” (Lenzerini, 2002). The second main component is Semantic Mapper or Transformation Agent. Semantic Mapper can be defined as tool that build the relation between the data sources and global schema by using mapping or matching techniques, there are many techniques to build the semantic mapper which will be discussed in. Within the semantic mapper, there are two main types of mapping. The first type of mapping is named as “global-as-view” (GAV). In GAV, “every element of the global schema is associated with a view over the sources” (Cali et al., 2005). The second type of mapping

is called as "local-as-view" (LAV). LAV "requires the sources to be defined as views over the global schema" (Cali et al. 2005). The main third component is the set of sources which can be defined as the raw resource of data that might be in different formats or types (Cali et al. 2005). A general formal framework for data integration was proposed by (Calì et al., 2005). Calì et al. (2005) used one of the logic from knowledge management and reasoning logic called as first-order logic (FOL) to construct the framework. Cali et al. (2005), introduced the general formal framework from theoretical general point of view without introducing a real-life solution. A semantic framework proposed by Zhu (2014) aims to improve data quality of electronic health records. Zhu (2014) by assessing the data quality introduced a new framework named "SemDQ". "SemDQ", designed by semantic web technologies as represented in OWL ontology language, the framework was constructed by Protégé 4.3 software developed by Stanford University. The global schema has been defined in the research as MOTeHR (MOT is the abbreviation of dataset name) which contain the international standards named open HER reference model which is available in XML format. The transformation agent has been developed to transform the results of SQL query into RDF datasets followed by transforming the XLS file to CSV file and add it to MOTeHR. Then data quality dimensions need to be defined by data quality criteria which contain SPARQL implementations.

Another ontology-based framework for electronic health records (eHR) has been proposed by González et al. (2011). While proposed framework by Zhu (2014) aims to improve the data quality by increasing data intrinsic; the proposed framework by (González et al., 2011) is looking to increase the data quality to improve data representation. The proposed framework by González et al. (2011) is aimed to integrate between three databases including patient, medical and laboratory record. The framework is architected by generic component model and the approach for mapping is common top-level ontology (González et al., 2011). The main idea for the framework architecture is to divide the integration process into four steps and each step linked with the step ontology González et al. (2011). The OWL used to build the global schema while the converter tool by different techniques used to transform multiple data format. Most of the semantic integration aims to enhance data quality by reducing data ambiguity nevertheless there is another approach increases the data ambiguity, the reason behind that is to hide the patient identity which corresponds to enhance the data quality accessibility by enhancing the data security and ensure data privacy. An example for that approach is a semantic framework proposed by (Martínez et al. 2013), in spite of the aforementioned semantic frameworks, the proposed framework uses a statistical method accompanied with a semantic framework to increase the privacy of medical published records by protecting it from any potential attack. Martínez et al. (2013) aim to achieve the target by increasing level of the non-numerical data indistinguishable, that drive Martínez et al. (2013) to use semantic integration (for non-numerical data) because the statistical algorithms can only work with numerical data.

Martínez et al. (2013) introduced three phases of the framework. The first phase is a comparison phase which measures the similarity and compares the terms with the medical knowledge base, then the second phase which is named as the aggregate phase is taking place by replacing several data records by one record. Finally, the third phase which is named as sorting phase finalizes the semantic process to prepare data for statistical processing. Martínez et al. (2013) used statistical disclosure method to complete the process after data has been adapted in previous phases. The methods used by Martínez et al. (2013), are as follows: the first method is recoding, which replaces the data attributes by another attribute. The second method is micro-aggregate, which uses maximum distance average vector method to generate data clusters. The third method is resampling, which select random sample, following be soring samples, then grouping and aggregation processed on records of each sample. We found proper assessment and evaluation for semantic integration framework proposed by Zhu (2014) and by Martínez et al. (2013), while both Zhu (2014) and by Martínez (2013), did not provide assessment or testing for a framework which was proposed by (González et al., 2011). The framework proposed by Zhu (2014) was assessed against three quality dimensions completeness, consistency, and timeliness by twelve criteria. The framework has been tested successfully on the simulated dataset to evaluate it (Zhu, 2014).

The framework proposed by González et al. (2011), has been evaluated by comparing between only statistical methods and statistical method with semantic framework in five cases. There is interrelation between semantic integration and other business domains; we reviewed published researches in telecommunication, financial and manufacturing domains which propose semantic integration framework to data and applications related to each domain. The proposed semantic framework by Wang (2008), which is related to manufacturing domain designed to integrate between multiple applications and dataset. Wang (2008) authors propose a conceptual semantic mediator to support semantic tasks to integrate multiple applications, the integration process run into phases:

First phase is to link each application or business process or data by web service then tag the web service and publish it in the semantic framework. This process is to identify the object to find out the data or business process need to be integrated. Then tag this data or business process based on sematic technology OWL. Finally, to publish it in SE-UDDI. The second phase is integration with collaborative applications, by analyzing the related data in the collaborative applications based on business requirements. Following with tagging it based on semantic technology then map the semantic data via mapper engine to SE-UDDI.

The proposed framework proposed by Wimmer et al. (2014) is related to integrate financial data from various resources internal and external. In the framework proposed by Wimmer et al. (2014), the semantic web technologies w3c is used to establish sematic integration between open linked data, XBRL files and public data. The ontology is design build based on financial interpretability ontology. The framework consists of two main component retrieval module and RDF generator. The retrieval module retrieve data from various resources of data by connecting through interfaces or agents, then the RDF generator is responsible to convert the retrieved data by using ontology based on financial interpretability ontology. The technology used to develop the RDF generator is SPARQL. The proposed framework by Wang (2008) is corresponding to the following data quality dimensions, that is completeness dimension, value added dimension and objectivity dimension, while the proposed framework by Wimmer et al. (2014) is corresponding to the following data quality dimensions, that is the value-added dimension, timeliness dimension and relevance dimension, these are the part of contextual data quality category (Wang & Strong 1996). Both of proposed frameworks Wang (2008) and Wimmer (2014) are conceptualize framework, although Wang (2008) and Wimmer (2014) apply the sematic framework on real life case, but the evaluation procedures or testing results are not shown. While the above-mentioned frameworks are conceptualized, the proposed sematic framework by Fuentes-Lorenzo et al. (2015) is based on business need for a company in telecommunication domain.

The proposed framework by Fuentes-Lorenzo et al. (2015) used Protégé language and semantic web technologies OWL and RDF to develop the framework. The framework consists of two main modules. The two main modules are mapping module and access module (Fuentes-Lorenzo et al. 2015). The mapping module contains two main components. The first one is internal data explicit mapping. The internal data is defined as the data carry on the meaning in itself. The second one is the explicit mapping. The explicit mapping has classes to classify and map sources to any corresponding class. Mapping between the class properties and source properties, mapping the object properties in direct or indirect relation, is included in explicit mapping. The selection of relation type is depended on number of resources between objects.

The access module contains three components basic query, advanced query and index search. The basic query provides end user with an ability to search for main resource or object while advanced query provides possibility to search with condition for resource or object. The index search provides the end use the ability to create a structured search.

The Fuentes-Lorenzo et al. (2015) framework enhances the data quality by realization of the following data quality dimensions, accuracy dimension, objectivity dimension which is under intrinsic category Wang & Strong (1996) and accessibility dimension which is under accessibility data quality category (Wang & Strong 1996). Krafft et al. (2010) proposed a framework for scientific domain. The solution named VIVO, which provide researchers a platform to share and integrate information and knowledge. The solution uses semantic integration framework by design map converter called RDF, SPARQL and using OWL technologies. The framework ontology is based on previous ontologies related to personal identification and bibliographic, the paper doesn't show the framework architecture.

2.3.3 Techniques

In the following section we will illustrate many techniques used for semantic integration, we will review traditional sematic integration techniques before we reviewed the latest sematic technologies.

2.3.3.1 RDF technology

The traditional technique for sematic integration was proposed Vdovjak & Houben (2001) by RDF technology. The technique is to develop a model for concepts and relationships, as an underlying model for the domain termed as "Conceptual module" (Vdovjak & Houben, 2001). The proposed framework by the proposed technique consists of five layers: source layer, XML instance layer, XML 2 RDF layer, Inference layer and application layer. The source layer consists of the different data sources like webpages or XML files as well as the RDF ontology. XML instance layer provides sequential XML, in XML 2 RDF layer each data source mapped to XML2RDF broker based on conceptual module. Inference layer that contain RDF mediator which consider the main component of the architecture. The mediator process

contains two steps, the first one finds out inference rules to apply it in inference engine, second distribute and decompose query to the brokers. The application layer can be any type of related applications like search agents.

2.3.3.2 Metadata ontology

One of the latest technique for semantic data integration proposed by Cverdelj-Fogaraši et al. (2017) is metadata ontology. The proposed technique is focussed on semantic integration for information systems. The technique is based, to provide semantics to the description of document metadata and enable sematic mapping between metadata of domain and metadata of another domain. The metadata ontology technique consists of three layers, the layers are service layer, data access layer and persistence layer. The technique uses the technology of ebXML, RIM standards to implement the metadata ontology, ebXML RIM standards is used to describe the metadata. The metadata ontology contains four components, which are core, classification, association and provenance.

Another latest technique for semantic data proposed by (Meng et al. 2017), the proposed new technique is based on crowdsourcing technology. The proposed technique is focus on sematic integration of knowledge bases. As current techniques of ontology of sematic integration for knowledge bases cluster data into classes as super class and sub-class, a problem arises for the current techniques of semantic integration which is related to taxonomy integration. Taxonomy integration is about how to find out the semantic relationships between same classes in different knowledge bases with different taxonomy, in case the classes is not equivalent. There are four types of entity matching: equivalent which can be defined as the both entities has the same concept, generalization which can be defined as one class in knowledge base considered as super class, while the same concept class in another knowledge base considered as sub-class, specification which can be defined as one class in knowledge base considered as sub-class while the corresponded concept class in the other knowledge base considered as super class, other which can be defined as which can be defined as the type of entities relationships not considered in the aforementioned three types.

Meng et al. (2017) proposed a technique based on crowd to resolve the taxonomy integration problem. Meng et al. (2017) faced two challenges while applying the introduced technique on knowledge base semantic integration: the first challenge is to keep the algorithm to perform like HIT (Human intelligence tasks-which defined as tasks done by human and the automated algorithms can't do). The solution was developed to resolve the challenge. The solution was named model local tree-based query, which consists of two components the query node and the local tree for the targeted node. The second challenge was to maximize the use of data block raised from crowd-sousing. The solution consists of two components, pruning power and utility function. There were two types of queries constructed into, static and adaptive query. The static query is based on utility function while the adaptive query is based on pruning power. The tested and evaluated technique on two real life knowledge bases was successful.

2.3.3.3 Semantic integration Technologies Challenges

Semantic integration considered as intersection between multiple technologies, many challenges arises accordingly, in this section we will discuss challenges of semantic technology from different point of views. One of the basic challenges for sematic integration is data heterogeneity; there are three types of data heterogeneity "Syntactic heterogeneity is caused by the use of different models or languages. Schematic heterogeneity results from structural differences. Semantic heterogeneity is caused by different meanings or interpretations of data in various contexts" (Cruz & Xiao, 2005). In addition to aforementioned challenge there are other challenges to implement the semantic integration architecture in real life (Doan et al., 2004).

Doan et al. (2004) discussed these challenges and figure it out into three main challenges: scalability of data: as most techniques does not evaluate on large amount of data, user interaction: as systems (from his point of view) cannot be work properly without human interference, mapping maintenance: require adaption for mapping from time to time to comply with changes in schema matching.

We will focus in this study on Sematic heterogeneity. A discussion for Sematic heterogeneity challenge was done by Doan et al. (2004), wherein the Sematic heterogeneity was analyzed and was found that the sematic heterogeneity is caused by two main reasons schema matching and entity matching. There are many reasons for schema matching challenge like difference in attributes, detail level or tagging or many other reasons. The reason for entity matching challenge is due to different naming for the entities like singer name and singer nickname as both are the same meaning for one entity. Schema matching has different type of matching like one to one or one to many or many to many which increases the challenge for schema matching.

Doan et al. (2004) introduce a solution for schema matching challenge by one of machine learning technique, by using similarity measure technique to find out the correspondence and rule based for similarity measure technique. The similarity measure technique was introduced by (Doan et al. 2005). To resolve schema matching challenge, different implementations of that technique were applied in applications like "CUPID" (Madhavan et al., 2001). Doan & Halevy (2005) favor similarity measure rule based more than other types of similarity measure because it is easy to use and faster than others. The other method is learning based which uses multiple approaches like neural network learning, Naïve Bayes and other learning-based approaches.

Another recent technique proposed by (De Carvalho et al., 2013) to handle Schema matching challenge, proposed approach based on one of artificial intelligence techniques named generational evolutionary algorithm. The algorithm proceeds with the following steps including initialization, evaluation, reproduction, selection, applying operation and presentation. De Carvalho et al. (2013) used two evaluation strategy to evaluate fitness function which are value oriented and entity oriented as well as cross over operation to swap between both components.

2.4 Data Quality Assessment

In this section we will discuss intersection between data quality assessment techniques and correlation with data quality improvement. It covers the impact of sematic integration to improve data quality (DQ). Firstly, we will discuss data quality assessment methods and concepts. Secondly, data quality assessment techniques (DQAT). Thirdly we will discuss the data quality assessment for specific domain. In the fourth part, we will discuss data quality assessment for linked data (LD) and lastly, we will discuss data quality assessment for rational databases.

2.4.1 Data quality assessment methods and concepts

To improve data quality there are two main strategies, the first strategy is improving the data value components and the second strategy is to improve the efficiency of processes to impact on the data quality at the end. The list of main seven techniques for first strategy (data-driven) has been presented by (Batini et al., 2009): Acquisition of new data, which improves data by acquiring higher-quality data to replace the values that raise quality problems. Standardization (or normalization), which replaces or complements nonstandard data values with corresponding values that comply with the standard. For example, nicknames are replaced with corresponding names, for example, Bob with Robert, and abbreviations are replaced with corresponding full names, for example, Channel Str. with Channel Street. Record linkage, which identifies those data representations in two or multiple tables that might refer to the same real-world object. Data and schema integration, which define a unified view of the data provided by heterogeneous data sources. Integration has the main purpose of allowing a user to access the data stored by heterogeneous data sources through a unified view of these data. Source trustworthiness selects data sources on the basis of the quality of their data. Error localization and correction identify and eliminate data quality errors by detecting the records that do not satisfy a given set of quality rules. Cost optimization, defines quality improvement actions along a set of dimensions by minimizing costs.

For the second strategy (process driven) two techniques has been formulated by (Batini et al., 2009): A reactive strategy is applied to data modification events, thus avoiding data degradation and error propagation. The reactive strategy is a resultant of Process control inserts checks and control procedures in the data production process when: new data are created, datasets are updated, new data sets are accessed by the process. Process redesign processes in order to remove the causes of poor quality and introduces new activities that produce data of higher quality.

2.4.2 Data quality assessment techniques

As high level of data quality is very important for business domain to achieve proper and high-quality information. Support, decision making or increase knowledge for specific subject, improvement of data quality cannot be achieved without data quality assessment. A data quality assessment provides the benchmark for measuring the quality of data.

However, the cost of data quality assessment to improve the data quality should be less than benefit from applying such technique, so according to that, we define one quality assessment technique named hybrid approach to be reviewed. Woodall et al. (2013) introduced a new technique for data quality assessment named hybrid approach; wherein the followed methodology developed the technique with a target to achieve the following quality goals that is validity, completeness, comprehension, understandability, test cover, practical utility and future resilience. A hybrid approach is aimed to combine data quality techniques to develop a new technique, the methodology started by finding out the

proper data quality assessment technique (DQAT) from available techniques. The condition was the DQAT should satisfy the criteria for full details provided and sufficient evaluation conducted. Then Woodall et al. (2013), cascaded the activities of each selected technique with suitable verification to avoid any missed activity, to assure achieving the quality goals authors present a conference paper to get feedback.

The development of data quality assessment technique included the following steps. The first step was to identify the objective of the assessment by finding out the targeted data quality problem and sort it according to importance. The second step is to identify the business domain needs for each data quality problem. The third step is to map the proper activity (gathered previously) to data quality problem, to fulfill the business domain need. The fourth step is to organize the mapped activities and dependences in an new data quality assessment technique (DQAT). The technique by Woodall et al. (2013) was designed for organization in maintenance, repair and operations domain, but we claim that the technique can be applied in various domains. The technique was applied on London underground to evaluate the applicability. The validity of the technique, according to trial test on London underground organization Woodall et al. (2013), found that the reference data was insufficient and require to be updated therefore Woodall et al. (2013) updated the technique.

2.4.3 Data quality assessment for specific domain

As data quality is critical for any domain, we select medical domain specifically electronic health records because data in the medical field is diversion data and dispersed as well as affect the human life.

Weiskopf et al. (2013) reviewed data quality dimension and quality assessments methods for electronic health records. Woodall et al. (2013) reviewed more than ninety research papers after an intensive selection process to find out the terms of data quality correspond to the main five data quality dimensions. Woodall et al. (2013) conclude finding terms into twenty-seven terms; then they map the finding terms with the five data quality dimensions. Woodall et al. (2013) conducted the same process to identify the most proper data quality assessment methods for electronic health records. The findings contain seven methods which can be categorized into three categories: the first category is comparison category which includes data source agreement, data element agreement and distribution comparison as these methods are based on comparison technique. Here data source comparison method compares between different data sources while data element agreement compare dataset with other data and distribution comparison. The distribution comparison compared between concepts and distribution of data summarized by statistics methods includes log review, gold standards, element presence and validity check. Second category is examination category contains the following methods, log review, element presence and validity check as these methods based on checking technique. The log review method is checking the data based on data attributes. The element presence method is to assure the proper data is present. The validity check method is using multiple mechanism to assure the reasonableness of data presented.

The third category is retrieval category which contains a gold standard method which is data driven from other sources without interventions. Weiskopf et al. (2013) dimension of the five-data quality dimension. The correlations between data quality dimensions and data quality assessment methods show high correlations between completeness, correctness and concordance dimensions and data quality assessment methods. While moderate correlation between plausibility and data quality assessment methods and correlation between currency and data quality assessment methods is low, as there is one data quality assessment method mapped to that dimension.

2.4.4 Data quality assessment for linked data

Linked data is increasing rapidly and integrate the datasets of linked data became very crucial. The high level of data quality is very important. So, data quality assessment for linked data is also becoming very important. This drive to establish frameworks and tools to integrate datasets of linked data. There are many frameworks and tools based on semantic integration proposed. In the following section, we will discuss one of linked data quality assessment framework and one tool for data quality assessment for linked data.

Mendes et al. (2012) proposed a framework named “sieve” which integrate with linked data integration framework. Linked data integration framework consists of three main component including web data access module, vocabulary mapping module and identity resolution module. Web data access module is responsible to import datasets using SPARQL technology. Vocabulary mapping module is responsible to map imported dataset using schema mapping. Finally, identity resolution module is responsible to find out similarities between entities from imported dataset based on Silk technology.

The framework composed from two modules including assessment module and fusion module (Mendes et al. 2012). The assessment module provides the user with a group of functions to help him to score DQ indicators as well as fusion function to resolve conflicts that arise from identity assessment module (Mendes et al. 2012). Fusion module provides user by a group of functions to deal with conflicts found by the previous module. Mendes et al. (2012) tested the proposed framework with real-life data and measured it with three data quality dimensions including completeness, consistency and conciseness, the evaluation showed high successful results (Mendes et al. 2012). To assess data quality of linked data, a tool named "TripleCheckMate" (Kontokostas et al., 2013) was introduced for linked data extracted from crowdsourcing process. The tools consist of two processes including manual process and semiautomatic process, the manual process is comprised of two steps. The first step is find out the quality problem and map it to the taxonomy of data problems. Second step is the evaluation step supported by proposed tool, the evaluation is based on predefined criteria of data quality to comprise data quality problems using crowdsourcing.

The tool supports the aforementioned processes and steps in the selection step. The tool provides the user with three selection options, then support the user to evaluate the resource triple by showing it to verify with the taxonomy. Provides the user the ability to extend the taxonomy in case of finding new type, if it does not exist earlier.

2.4.5 Data quality assessment for rational databases

Multiple and different types of databases for information systems of organizations is becoming one of major challenges, as there are systems for enterprise information systems like enterprise resource planning systems, systems for supply chain management, a system for inventory management like systems use RFID technologies, material requirement planning and many other types of systems. Traditional integration is not sufficient to fulfill different business domain requirements and need for more integration tool arises. The reason for the need for more advanced methodologies for integration is due to the lack of traditional integration to meet the data quality objectives. In this section, we will discuss data quality improvement by using semantic integration. Firstly, we will discuss an architecture, to improve data quality for the cooperative information system. Secondly, we will discuss new method to improve data quality. A new framework named "DaQuinCIS" Scannapieco et al. (2004) has been introduced for cooperative information systems. The architecture is based on four data quality dimensions. Dimensions provide the proper data quality assessment with subsequently improving data quality. The four data quality dimensions are consistency, accuracy, currency and completeness. The architecture consists of four main components including rating service, data quality broker, quality factory, quality notification service.

Each dataset based on the data quality dimensions mentioned above is being evaluated by quality factory then by group of functions. The quality data broker module retrieves selected data from different data sources. The module use peer to peer methodology which provides more flexibility and use global as view approach to process queries then the rating service. The rating service is a third-party application used to validate data from different data sources. Finally, data notification service acts like messenger between services and data source, to inform about data availability or changes in data quality. The architecture is mainly used as a model for data retrieval named "Data Quality model" (Scannapieco et al., 2004). Data Quality model contains quality types which represent data quality values and map it to data quality dimensions. Here, the model schema is transformed to XML schema. The model schema consists of three types of data that is: schema of each retrieved data, the schema for quality data and relation between both. New method to improve data quality using semantic integration with reference to databases, has been discussed in the research paper introduced by (Brouwer et al., 2016) which consists of the following steps: assess data quality of individual datasets, create a shared ontology, Perform Semantic matching, Perform Semantic integration, Evaluate data quality improvement. Assess data quality step is based on assessing the dataset quality by data quality dimensions which are categorized into four categories proposed by Wang (1996). Create a shared ontology is processed in three phases that is the ontology capture, ontology coding and integrating existing ontologies. In ontology capture phase author identify the definitions, concepts, terms, and structure required for the ontology. In ontology coding phase author concentrate to model the meta-ontology data. In integrate existing ontologies phase author integrate prior ontologies create before the project start. The semantic matching step is constructed by one to one relationship. The semantic integration step is to combine the data sources based on semantic matching into one dataset. The evaluate data quality improvement step determines whether the data quality has been improved or not. The above method to improve data quality was

evaluated in real life dataset by integrating two real-life datasets related to improving quality of carbon footprint data. The evaluation result shows that method is implemented successfully.

3 Methodology and Findings

We found rich of articles in bioinformatics and health information systems within literature review phase as out of one hundred thirty-six articles screened. We found eighty-seven articles related to bioinformatics and health information systems with respect to semantic integration technologies or assessment to data quality.

3.1 Comparison between selected semantic integration frameworks

We compared selected frameworks discussed earlier in literature review section based on the comparison criteria discussed earlier in the research methodology chapter. Based on comparison table 2, we started selection process to find qualified framework by excluding conceptual approach frameworks, and then exclude frameworks not able to apply for proposed framework. Finally, three frameworks were left, we select framework "SemDQ" proposed by (Zhu, 2014). Zhu (2014) as it fulfills the highest number of data quality dimensions.

Study	Approach	Release Date	Qualify to apply	No of Data Quality Dimensions	Applied Data Quality Dimension
(Calì et al., 2005)	Conceptual	2005	No	1	completeness
(Zhu, 2014)	Practical	2014	Yes	3	completeness, consistency, accuracy
(González et al., 2011)	conceptual	2011	No	3	completeness, consistency, timeliness
(Martínez et al., 2013)	Practical	2013	Yes	2	completeness, consistency
(Wang, 2008)	conceptual	2008	Yes	2	completeness, consistency
(Wimmer et al., 2014)	conceptual	2014	Yes	3	completeness, consistency, timeliness
(Fuentes-Lorenzo et al., 2015)	Practical	2015	Yes	2	completeness, consistency
(Krafft et al., 2010)	Practical	2010	No	1	completeness

Table 2. Comparison between selected semantic integration frameworks.

3.2 Comparison between selected semantic integration approaches

We used the same selection methodology to select the proper approach, as all approaches are conceptual; we start selection process by excluding the oldest approaches, then exclude the approach cannot be applied for Enterprise Information systems, so the selected approach is a declarative semantics approach proposed by (Caroprese & Zumpano, 2017). Table 3 summarizes the comparison results.

Proposed By	Approach	Release Date	Qualify to apply	No of Dimension	Applied Dimension
(Calvanese et al. 2003)	conceptual	2003	No	2	completeness, timeliness
(Cruz et al. 2004)	conceptual	2004	Yes	1	completeness
(Dimartino et al. 2015)	conceptual	2015	No	2	completeness, timeliness
(Caroprese et al. 2017)	conceptual	2017	Yes	2	completeness, consistency

Table 3. Comparison between selected semantic integration approaches.

3.3 Comparison between selected semantic integration Techniques

We apply the same selection methodology to select the qualified technique; at the beginning, we exclude technique that not able to qualify. We found the other two techniques are equal in the remaining

other selection criteria expected of release date. However, we selected both techniques as both will match we data sources required for the proposed hybrid framework. A technique proposed by Meng et al. (2017), will support the proposed framework to integrate document management systems, and the technique proposed by (Vdovjak & Houben, 2001) will support the proposed framework to integrate XML files. Table 4 summarizes the results of the comparison.

Proposed By	Approach	Release Date	Qualify to apply	No of Dimension	Applied Dimension
(Vdovjak, & Houben 2001)	Conceptual	2001	Yes	2	completeness, timeliness
(Cverdelj-Fogaraši et al. 2017)	Practical	2017	Yes	2	completeness, timeliness
(Meng et al. 2017)	Practical	2017	No	2	completeness, timeliness

Table 4. Comparison between selected semantic integration Techniques.

3.4 The Proposed Hybrid Methodology

The proposed methodology is based on peer to peer approach for deductive databases which provide the methodology flexibility to use additional data sources without the need to update the global schema.

3.4.1 Challenges and Conflicts

We face the following challenges while composing the framework:

- Map each selected component to the right place in the proposed framework and map each data source type to proper layer.
- To adapt the component to other proposed framework components.

There is conflict discovered between the qualified framework and RDF technique for the ontology of the core approach component. We fix the conflict by eliminate the RDF technique ontology and merge it into framework ontology and map knowledge bases to it.

3.4.2 Framework components

The proposed framework consists of six layers and one component. Each layer and component perform as follows:

- Source layer: contain the data source from different data types, database, spreadsheet files, web sources and document management system.
- Translation Layer is composed of two components.
- Document Metadata Translator is responsible for translating unstructured data from Document management system into RDF triple store using the crowdsourcing technique. By processing the document into three layers, that is service layer, data access layer and persistence layer.
- Transformation Agent is responsible for transforming database dataset to RDF dataset by using SQL statements and convert spreadsheet files to CSV files then transform it to RDF dataset.

XML layer is responsible for wrapping XML files and serializing it.

- RDF layer consists of RDF brokers which is responsible for preparing RDF dataset files as RDF require some conventions to be in proper RDF interpretation to run sub-queries received from the mediator.
- Inference Layer contains the core component which is the mediator, as it contains the rules, classes, properties, and mapping between classes, properties and RDF parser. The mediator is responsible for decomposing the main query to sub-queries, forward it to RDF brokers and apply rules on returned data from brokers.
- Application layer contains various types of application which send the main query to a mediator or another type of application with both ways connection. The queries can be sent to the mediator and receive data for integration purpose.
- Ontology Component provides the methodology by ontology required for a mediator; it contains structures, concepts, and classes. The ontology is interlinked with the internal knowledge base and external knowledge base to update the ontology classes and concepts.

The proposed hybrid semantic integration methodology is depicted in Figure 2.

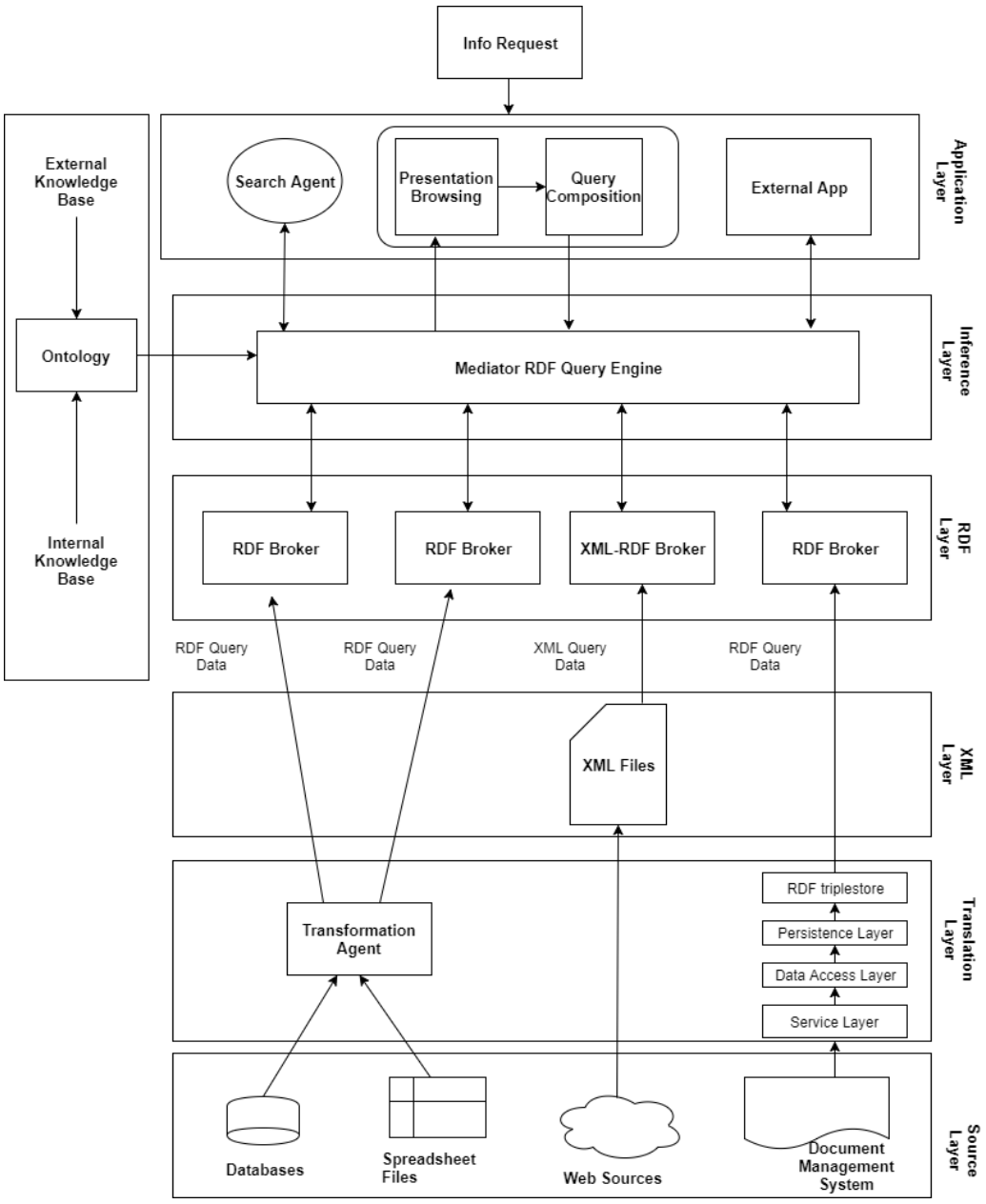

Figure 2. The proposed Hybrid Methodology for Semantic Integration.

3.4.3 Framework Process Flow

Based on bottom-up process flow approach, we demonstrate data flow through framework as follows:

- Dataset from database and spreadsheet files flows from the source layer to transformation agent while data from DMS flow to RDF translator process in translation layer, while dataset from web sources flows directly to XML layers.
- Specific query data run on each transformed data as RDF query applied on transformed data received from databases, spreadsheet files, and DMS wile XML query applied on transformed data received from web sources.
- Brokers in RDF layer receive output from query data and map it with sub-query received from the mediator.

- Mediator RDF Query Engine in Inference layer is the core component, receives an initial query from the application layer and decompose it to sub-query forwarded to brokers, Mediator RDF Query Engine to apply rules to map between data entities based on ontology received from ontology component.
- Ontology component consists of ontology element integrated with the internal & external knowledge base

Application Layer receives the info request, forward the initial query to the mediator and receives results; and finally, forward the results to the requester.

3.4.4 Data Quality Measurement

- The proposed framework will improve data quality by improving the level of the following data quality dimensions:
- Completeness: Data retrieved from each data source complete each other. For instance, data received from enterprise resource planning system, with data received from user spreadsheet files and the supporting documents from document management system will complete each other.
- Consistency: information retrieved is consistent regardless of the format of different data sources.
- Accuracy: Avoid human mistake risk and improve dataset accuracy from a data source by validating it with a dataset from another data source.
- For instance, integrate the exchange rate from web source with overseas vendor invoice will provide user with exact local currency.
- Timeliness: by automating the process, all required information gathered to the user is in very limited time.

4 Conclusion

In this study, we aim to improve the level of data quality by applying new semantic integration framework. To achieve our target, we screened one hundred thirty-six research articles and two books. Afterword, we select thirty-five articles and the two books based on specific criteria. Finally, we classified and categorized them based on the related research question, technology, and business domain. We reviewed the selected articles then we did a comparative study to select the qualified technologies for the proposed framework. The new hybrid framework consists of six layers and one competent. The six layers are source layer, a translation layer, XML layer, RDF layer, inference layer, application layer, and an ontology component. During the process of integration and development of the proposed framework, we fixed mapping issues and adaptation challenges as well as ontology conflicts. Finally, we verified the new proposed hybrid framework for semantic integration on data quality dimensions. Our findings showed that the proposed framework was successful to achieve the following data quality dimensions: completeness, consistency, accuracy, and timeliness.

References

Al Emran, M., & Shaalan, K. (2014, September). A survey of intelligent language tutoring systems. In *Advances in Computing, Communications and Informatics (ICACCI, 2014 International Conference on* (pp. 393-399). IEEE.

Al-Emran, M. (2015). Hierarchical Reinforcement Learning: A Survey. *International Journal of Computing and Digital Systems*, *4*(2).

Al-Emran, M., & Malik, S. I. (2016). The impact of google apps at work: higher educational perspective. *International Journal of Interactive Mobile Technologies (iJIM)*, *10*(4), 85-88.

Al-Emran, M., Zaza, S., & Shaalan, K. (2015, May). Parsing modern standard Arabic using Treebank resources. In *Information and Communication Technology Research (ICTRC), 2015 International Conference on* (pp. 80-83). IEEE.

Al-Qaysi, N., & Al-Emran, M. (2017). Code-switching Usage in Social Media: A Case Study from Oman. *International Journal of Information Technology and Language Studies*, *1*(1), 25-38.

Askham, N., Cook, D., Doyle, M., Fereday, H., Gibson, M., Landbeck, U., ... & Schwarzenbach, J. (2013). The six primary dimensions for data quality assessment. Technical report, DAMA UK Working Group.

Batini, C., Cappiello, C., Francalanci, C., & Maurino, A. (2009). Methodologies for data quality assessment and improvement. ACM computing surveys (CSUR), 41(3), 16.

Brouwer, O. F. (2016). Applying Semantic Integration to improve Data Quality (Master's thesis).

Calì, A., Lembo, D., & Rosati, R. (2005). A comprehensive semantic framework for data integration systems. Journal of Applied Logic, 3(2), 308-328.

Calvanese, D., Damaggio, E., De Giacomo, G., Lenzerini, M., & Rosati, R. (2003, September). Semantic data integration in P2P systems. In International Workshop on Databases, Information Systems, and Peer-to-Peer Computing (pp. 77-90). Springer Berlin Heidelberg.

Caroprese, L., & Zumpano, E. (2017, August). A Declarative Semantics for P2P Systems. In International Cross-Domain Conference for Machine Learning and Knowledge Extraction (pp. 315-329). Springer, Cham.

Cruz, I. F., & Xiao, H. (2005). The role of ontologies in data integration. Engineering intelligent systems for electrical engineering and communications, 13(4), 245.

Cruz, I. F., Xiao, H., & Hsu, F. (2004, July). Peer-to-peer semantic integration of XML and RDF data sources. In AP2PC (Vol. 3601, pp. 108-119).

Cverdelj-Fogaraši, I., Sladić, G., Gostojić, S., Segedinac, M., & Milosavljević, B. (2017). Semantic integration of enterprise information systems using meta-metadata ontology. Information Systems and e-Business Management, 15(2), 257-304.

De Carvalho, M. G., Laender, A. H., GonçAlves, M. A., & Da Silva, A. S. (2013). An evolutionary approach to complex schema matching. Information Systems, 38(3), 302-316.

Dimartino, M. M., Calì, A., Poulovassilis, A., & Wood, P. T. (2015). Peer-to-peer semantic integration of linked data. In CEUR Workshop Proceedings (Vol. 1330, pp. 213-220). CEUR Workshop Proceedings.

Doan, A., & Halevy, A. Y. (2005). Semantic integration research in the database community: A brief survey. AI magazine, 26(1), 83.

Doan, A., Noy, N. F., & Halevy, A. Y. (2004). Introduction to the special issue on semantic integration. ACM Sigmod Record, 33(4), 11-13.

Fuentes-Lorenzo, D., Sánchez, L., Cuadra, A., & Cutanda, M. (2015). A RESTful and semantic framework for data integration. Software: Practice and Experience, 45(9), 1161-1188.

Fürber, C. (2015). Data quality management with semantic technologies. Springer.

Goh, C. H. (1997). Representing and reasoning about semantic conflicts in heterogeneous information systems (Doctoral dissertation, Massachusetts Institute of Technology).

González, C., Blobel, B. G., & López, D. M. (2011). Ontology-based framework for electronic health records interoperability. Studies in health technology and informatics, 169, 694-698.

Gruber, T. (1993). What is an Ontology. WWW Site http://www-ksl. stanford. edu/kst/whatis-an-ontology. html (accessed on 07-09-2004).

Guido, A. L., & Paiano, R. (2010). Semantic integration of information systems. International Journal of Computer Networks and Communications (IJCNC), 2.

Hammer, M., & Champy, J. (2009). Reengineering the Corporation: Manifesto for Business Revolution, A. Zondervan.

Izza, S. (2009). Integration of industrial information systems: from syntactic to semantic integration approaches. Enterprise Information Systems, 3(1), 1-57.

Khoei, A. (2013). Towards an ontology for data quality in integrated chronic disease management: a realist review of the literature. International journal of medical informatics, 82(1), 10-24.

Kontokostas, D., Zaveri, A., Auer, S., & Lehmann, J. (2013, October).

Krafft, D. B., Cappadona, N. A., Caruso, B., Corson-Rikert, J., Devare, M., Lowe, B. J., & VIVO Collaboration. (2010). Vivo: Enabling national networking of scientists.

Lenzerini, M. (2002, June). Data integration: A theoretical perspective. In Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems (pp. 233-246). ACM.

Liaw, S. T., Rahimi, A., Ray, P., Taggart, J., Dennis, S., de Lusignan, S. & Talaei-

Madhavan, J., Bernstein, P. A., & Rahm, E. (2001, September). Generic schema matching with cupid. In vldb(Vol. 1, pp. 49-58).

Martínez, S., Sánchez, D., & Valls, A. (2013). A semantic framework to protect the privacy of electronic health records with non-numerical attributes. Journal of Biomedical Informatics, 46(2), 294-303

Mendes, P. N., Mühleisen, H., & Bizer, C. (2012, March). Sieve: linked data quality assessment and fusion. In Proceedings of the 2012 Joint EDBT/ICDT Workshops (pp. 116-123). ACM.

Meng, R., Chen, L., Tong, Y., & Zhang, C. (2017). Knowledge Base Semantic Integration Using Crowdsourcing. IEEE Transactions on Knowledge and Data Engineering, 29(5), 1087-1100.

Mhamdi, C., Al-Emran, M., & Salloum, S. A. (2018). Text Mining and Analytics: A Case Study from News Channels Posts on Facebook. In *Intelligent Natural Language Processing: Trends and Applications* (pp. 399-415). Springer, Cham.

Prabhakaran, M., & Chou, C. (2006). Semantic Integration in Enterprise Information Management. SETLabs, 4(2), 45-52.

Salloum, S. A., Al-Emran, M., & Shaalan, K. (2016). A survey of lexical functional grammar in the Arabic context. *Int. J. Com. Net. Tech*, *4*(3), 430.

Salloum, S. A., Al-Emran, M., & Shaalan, K. (2017a). Mining Text in News Channels: A Case Study from Facebook. *International Journal of Information Technology and Language Studies*, *1*(1), 1-9.

Salloum, S. A., Al-Emran, M., & Shaalan, K. (2017b, March). Mining social media text: extracting knowledge from facebook. In *International Journal of Computing and Digital Systems* (Vol. 6, No. 2). University of Bahrain.

Salloum, S. A., Al-Emran, M., Abdallah, S., & Shaalan, K. (2017c, September). Analyzing the Arab Gulf Newspapers Using Text Mining Techniques. In *International Conference on Advanced Intelligent Systems and Informatics* (pp. 396-405). Springer, Cham.

Salloum, S. A., Al-Emran, M., Monem, A. A., & Shaalan, K. (2017d). A survey of text mining in social media: facebook and twitter perspectives. *Adv. Sci. Technol. Eng. Syst. J*, *2*(1), 127-133.

Salloum, S. A., Al-Emran, M., Monem, A. A., & Shaalan, K. (2018a). Using Text Mining Techniques for Extracting Information from Research Articles. In *Intelligent Natural Language Processing: Trends and Applications* (pp. 373-397). Springer, Cham.

Salloum, S. A., AlHamad, A. Q., Al-Emran, M., & Shaalan, K. (2018b). A Survey of Arabic Text Mining. In *Intelligent Natural Language Processing: Trends and Applications* (pp. 417-431). Springer, Cham.

Salloum, S. A., Mhamdi, C., Al-Emran, M., & Shaalan, K. (2017e). Analysis and Classification of Arabic Newspapers' Facebook Pages using Text Mining Techniques. *International Journal of Information Technology*, *1*(2), 8-17.

Scannapieco, M., Virgillito, A., Marchetti, C., Mecella, M., & Baldoni, R. (2004). The DaQuinCIS architecture: a platform for exchanging and improving data quality in cooperative information systems. Information systems, 29(7), 551-582

Shan, T. C. & Hua, W. W. (2006). Taxonomy of java web application frameworks. In E-business engineering, 2006. icebe'06. ieee international conference on (pp. 378–385). IEEE.

Shvaiko, P., & Euzenat, J. (2013). Ontology matching: state of the art and future challenges. IEEE Transactions on knowledge and data engineering, 25(1), 158-176.

Triplecheckmate: A tool for crowdsourcing the quality assessment of linked data. In International Conference on Knowledge Engineering and the Semantic Web (pp. 265-272). Springer, Berlin, Heidelberg.

Vdovjak, R., & Houben, G. J. (2001, April). RDF-Based Architecture for Semantic Integration of Heterogeneous Information Sources. In Workshop on information integration on the Web (pp. 51-57).

Wand, Y., & Wang, R. Y. (1996). Anchoring data quality dimensions in ontological foundations. Communications of the ACM, 39(11), 86-95.

Wang, Q. (2008, December). Semantic framework model-based intelligent information system integration mode for Manufacturing Enterprises. In Intelligent Information Technology Application, 2008. IITA'08. Second International Symposium on (Vol. 1, pp. 223-227). IEEE.

Wang, R. Y., & Strong, D. M. (1996). Beyond accuracy: What data quality means to data consumers. Journal of management information systems, 5-33

Weiskopf, N. G., & Weng, C. (2013). Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. Journal of the American Medical Informatics Association, 20(1), 144-151.

Wimmer, H. & Narock, T. (2014). "Integrating and improving quality, trust, and comprehension of financial data using Semantic Web Technologies: A proposed conceptual model." Northeast Decision Sciences Annual Conference, Philadelphia, Pennsylvania 2014.

Woodall, P., Borek, A., & Parlikad, A. K. (2013). Data quality assessment: the hybrid approach. Information & management, 50(7), 369-382.

Zaveri, A., Rula, A., Maurino, A., Pietrobon, R., Lehmann, J., & Auer, S. (2016). Quality assessment for linked data: A survey. Semantic Web, 7(1), 63-93.

Zaza, S., & Al-Emran, M. (2015, October). Mining and exploration of credit cards data in UAE. In *e-Learning (econf), 2015 Fifth International Conference on* (pp. 275-279). IEEE.

Zhu, L. (2014). *SemDQ: A Semantic Framework for Data Quality Assessment* (Master's thesis, University of Waterloo).