Big Research Data Integration

Valentina Bartalesi, Carlo Meghini, and Costantino Thanos

Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo" – CNR, Pisa, Italy
{valentina.bartalesi,carlo.meghini,costantino.thanos}@isti.cnr.it

Abstract. The paper presents a vision about a new paradigm of data integration in the context of the scientific world, where data integration is instrumental in exploratory studies carried out by research teams. It briefly overviews the technological challenges to be faced in order to successfully carry out the traditional approach to data integration. Then, three important application scenarios are described in terms of their main characteristics that heavily influence the data integration process. The first application scenario is characterized by the need of large enterprises to combine information from a variety of heterogeneous data sets developed autonomously, managed and maintained independently from the others in the enterprises. The second application scenario is characterized by the need of many organizations to combine information from a large number of data sets dynamically created, distributed worldwide and available on the Web. The third application scenario is characterized by the need of scientists and researchers to connect each others research data as new insight is revealed by connections between diverse research data sets. The paper highlights the fact that the characteristics of the second and third application scenarios make unfeasible the traditional approach to data integration, i.e., the design of a global schema and mappings between the local schemata and the global schema. The focus of the paper is on the data integration problem in the context of the third application scenario. A new paradigm of data integration is proposed based on the emerging new empiricist scientific method, i.e., data driven research and the new data seeking paradigm, i.e., data exploration. Finally, a generic scientific application scenario is presented for the purpose of better illustrating the new data integration paradigm, and a concise list of actions that must be performed in order to successfully carry out the new paradigm of big research data integration is described.

Keywords: Research Data Integration · Big Data · Semantic Web

1 Introduction

Data Integration has the goal of enriching and completing the information available to the users by adding complementary information residing at diverse information sources [24,12]. It aims at providing a more comprehensive information

basis in order to better satisfy user information needs. This is achieved by combining data residing at diverse data sets and creating a unified view of these datasets. This view provides a single access point to these distributed, heterogeneous and autonomous data sets. Therefore, it frees the user from the necessity of interacting separately with each of these data sets. We distinguish two types of data integration [3]. The first type of data integration, structural data integration, refers to the ability to accommodate in a common data representation model distributed data sets represented in different data representation models and formats. In essence, in this type of integration the goal is to augment the dimensionality of an entity/object represented in different distributed data sets by collecting together all the attributes/features associated with this entity/object. The second type of data integration, semantic data integration [8], refers to the ability to combine distributed data sets on the basis of existing semantic relationships between them. In essence, in this type of integration the goal is to augment the relationality of an entity/object represented in a data set by linking it to entities/objects semantically closely related to it and represented in other distributed data sets.

The type of data integration very much depends on the characteristics of the data to be integrated. In the scientific domain, data can be referred to as raw or derivative. Raw data consist of original observations, such as those collected by satellite and beamed back to earth or generated by an instrument or sensor or collected by conducting an experiment. Derivative data are generated by processing activities. The raw data are frequently subject to subsequent stages of curation and analysis, depending on the research objectives. While the raw data may be the most complete form, derivative data may be more readily usable by others as processing usually makes data more usable, thus increasing their intelligibility. Structural data integration is, mainly, performed between data sets containing raw data, while semantic data integration is more appropriate for data sets containing derivative data; in this case, the semantic relationship between data sets is, usually, a correlation between them.

We have identified three main application scenarios where data integration is of paramount importance [1]. These three application scenarios well illustrate the evolution of the data integration concept both from the application and technological point of view. The first application scenario is characterized by the need of large enterprises to combine information from a variety of heterogeneous data sets developed autonomously, managed and maintained independently from the others in the enterprises. The second application scenario is characterized by the need of many organizations to combine information from a large number of data sets dynamically created, distributed worldwide and available on the Web. The third application scenario is characterized by the need of scientists and researchers to connect each other's research data as new insight is revealed by connections between diverse research data sets. In essence, we can say that data integration, in the first application scenario, is instrumental in effectively and efficiently managing large enterprises and in supporting the enterprise' planning activities. In the second application scenario, data integration is, mainly,

instrumental in activities like data mining, forecasting, statistical analysis, decision making, implementing strategy, etc., conducted by organizations whose business is based on the analysis and comparison of data stored in a large number of data collections distributed worldwide. In the third application scenario, data integration is, mainly, instrumental in exploratory studies carried out by research teams. For each one of the above three application scenarios, different technological challenges must be faced in order to develop integration systems that efficiently and effectively carry out the data integration process.

The paper is organized as follows: in Section 2, it overviews the technological challenges to be faced in order to successfully carry out the traditional data integration process. These challenges have been extensively discussed in the literature. In Section 3, the features of the three application scenarios that heavily influence the data integration process are described. In Section 4 a new paradigm of big research data integration is described. This is the main contribution of the paper. In 4 Subsections the enabling technologies are identified and briefly described. In Section 5 a generic scientific application scenario is presented for the purpose of better illustrating the new data integration paradigm. Finally, in Section 6, a concise list of actions that must be performed in order to successfully carry out the new paradigm of big research data integration is described.

2 Data Integration Technological Challenges

The traditional approach to data integration is a three-step process: data transformation, duplicate detection, and data fusion [19]. Data transformation is concerned with the transformation of the local data representations (local schemata) into a common representation, the global schema or mediated schema, which hides the structural aspects of the different local data collections. Two basic approaches have been proposed for this purpose [17]. The first approach, called Global-as-View (GAV), requires the global schema be expressed in terms of the local data schemata. In essence, this approach regards the local data schemata and generates a global schema that is complete and correct with respect to the local data schemata, and is also minimal and understandable. The second approach, called Local-as-View (LAV), requires the global schema be specified independently from the local data schemata, and the relationships between the global schema and the local data schemata are established by defining every local data schema as a view over the global schema.

The global schema can be materialized or virtual. In the first case, it is materialized in a persistent store, for example, in a data warehouse that consolidates data from multiple data sets. Extract-Transform-Load (ETL) tools [22] are used to extract, to transform, and load data from several data sets into a data warehouse. In the second case, the global schema is not materialized; it is virtual, that is, it just gives the illusion that the data sets have been integrated. The users pose queries against this virtual global schema; a mediator, i.e., a software module, translates these queries into queries against the single data sets and integrates the result of them. The second step, i.e., duplicate detection, is con-

cerned with the identification of multiple, possibly, inconsistent representations of the same real-world entities. The third step, i.e., data fusion [19], is concerned with the fusion of the duplicate representations into a single representation and the resolution of the inconsistencies in the data.

Several technological challenges must be faced in order to efficiently and effectively carry out the data integration process. The first challenge to be faced regards the structural heterogeneity of the different data sets to be integrated. Integrating different data models and formats, i.e., structural data integration, requires the resolution of several types of conflicts. At the schema level, different data schemas may use (i) different data representations; (ii) different scales and measurement units; and (iii) different modelling choices, for example, an entity in one schema is represented as an attribute in another schema. At the data level, some contradictions may occur when different values exist for an attribute of the same entity. In addition, uncertainty may occur when a value of an attribute is missing in one data collection and is present in another data collection.

The second and more demanding challenge to be faced regards the semantic heterogeneity of the different data schemas to be integrated, i.e., semantic data integration. In general, the schemas of the different data sets do not provide explicit and precise semantics of the data to be integrated. The lack of precise data semantics can cause semantic ambiguities. For example, it may occur that two relations in two different data sets (assuming that both collections are modeled according to the relational data model) have the same name but heterogeneous semantics. This can induce in making erroneous design choices when the mediated schema is designed. In addition, the meanings of names and values may change over time. To mitigate the problem of semantic heterogeneity data must be endowed with appropriate metadata.

The third challenge to be faced regards the implementation of a mediating environment that provides a core set of intermediary services between the global schema and the local schemata. Such core set of services should include services that [23]:

- quickly and accurately find data that support specific user needs;
- map data structures, properties, and relationships from one data representation
- scheme to another one, equivalent from the semantic point of view;
- verify whether two strings/patterns match or whether semantically heterogeneous data match;
- optimize access strategies to provide small response time or low cost;
- resolve domain terminology and ontology differences; and
- prune data ranked low in quality or relevance.

In essence, the intermediary services must translate languages, data structures, logical representations, and concepts between global and local schemata. The effectiveness, efficiency, and computational complexity of the intermediary functions very much depend on the characteristics of the information models (expressiveness, semantic completeness, adequate modelling of data descriptive

information (metadata), reasoning mechanisms, etc.) and languages adopted by the user. Ideally, they must provide a framework for semantics and reasoning. An important component of the mediating environment is ontologies [10]. Several domain-specific ontologies are being developed (gene ontology, sequence ontology, cell type ontology, biomedical ontology, CIDOC, etc.). Ontologies have been extensively used to support all the intermediary functions because they provide an explicit and machine-understandable conceptualization of a domain.

3 The Main Characteristics of the Three Application Scenarios

The first application scenario is characterized by:

- a relatively small number of data sets to be integrated;
- relatively static local schemas;
- fixed local schemas; and
- local schemas known in advance.

In such scenario, the design of the global schema as well as the mappings between the local schemas and the global schema are relatively easy tasks. Several data integration systems have been implemented which operate efficiently in this scenario [24,5].

The second application scenario is characterized by:

- large number of data sets to be integrated;
- data sets containing huge volumes of data;
- data sets of widely differing data qualities;
- extremely heterogeneous local schemas;
- dynamically created local schemas (not fixed); and
- local schemas not known in advance.

In such an application environment, performing the data integration process is a very difficult task [9]. In fact:

- the large number of data sets to be integrated makes the alignment of their schemas very difficult;
- the extreme heterogeneity of data representation models adopted by the data sets to be integrated makes the design of a global schema very difficult;
- the dynamicity of the local schemas requires that the global schema undergoes continuous changes/extensions;
- the dynamicity of the local schemas makes difficult understanding the evolution of semantics and infeasible the capture of data changes timely;
- the widely differing qualities of the data to be integrated makes the alignment of the local schemas really hard; and
- the large volumes of the data to be integrated makes their warehousing very expensive.

In order to overcome the difficult problem of designing and maintaining a global schema in such a complex and dynamic context, it has been proposed to adopt an ontology-based approach to data management [7]. In this approach, the global schema is replaced by the conceptual model of the application domain. Such a model is formulated as an ontology expressed in a logic-based language. With this approach the integrated view is a semantically rich description of the relevant concepts in the domain of interest, as well as the relationships between such concepts. The users of an ontology-based data management system are enabled to query the data using the elements in the ontology as predicates. The ontology-based approach permits to overcome nicely the need for continuously reshaping of the global schema as an ontology can be easily extended. In fact, this approach supports an incremental process in representing the application domain. The domain ontology can be enriched with new concepts and relationships between them as new data sources or new elements in these sources are added. In essence, this approach supports the evolution of the ontology and the mappings between the ontology concepts and the data contained in the data sources supporting, thus, a "pay-as-you-go" data integration.

A conceptual difference between the above two data integration application scenarios regards the global perspective to be taken into account when designing global schemata. In the first application scenario, the designer of the global schema, by adopting the LAV approach, is enabled to take into due consideration a global perspective concerning the enterprise's activities. In the second application scenario, such an opportunity (i.e., the LAV approach) is not possible as the design of a global schema is practically unfeasible.

The third application scenario is situated within the scientific world. Indeed, we focused on the characteristics that are specific to data produced by research activities in the context of a new scientific framework characterized by; (i) the production of big data; and (ii) a science increasingly data intensive, multidisciplinary and e-science. In this world, some disciplines, for example, astronomy and high-energy physics, rely on a limited number of data repositories containing huge amounts of data. In that situation, researchers know where to find data of interest for their research activities. The problem is that the amount of data contained in these repositories outgrows the capabilities of query processing technology. In the case of overwhelming amounts of data, a new paradigm of query processing has been proposed: "data exploration" [14]. This new paradigm enables us to re-formulate the data integration problem as, mainly, a data interconnection problem. Exploration-based systems, instead of considering a huge data set in one go, incrementally and adaptively guide users towards the path that their queries and the result lead. These systems do not offer a correct and complete answer but rather a hint of how the data looks like and how to proceed further, i.e., what the next query should be. Data exploration, therefore, is a new approach in discovering connections and correlations between data in the big data era. Some other disciplines rely on large number of voluminous data sets with varying representation models, formats and semantics produced by many Labs and research groups distributed worldwide. Often, data of the same phe-

nomenon come from many data sets. In such an application environment, discovering connections and correlations between data from autonomous distributed data sets is driving the need for data integration [6]. Integrating multiple data sets will enable science to advance more rapidly and in areas heretofore outside the realm of possibility. The third application scenario we consider addresses exactly the need for data integration of these scientific disciplines. Such an application scenario shares all the main characteristics of the second application scenario; in addition, it has some peculiar characteristics:

- the local schemata continuously evolve as new insights are gained in a scientific domain; for example, certain concepts can be invalidated in the light of new discoveries.
- data heterogeneity that is also created by the fact that researchers can conceptualize the same scientific problem in different ways due to the fact that, for example, belong to different "schools of thought".
- data heterogeneity that is intrinsic to some scientific disciplines. In fact, as reported in [16]: "in the environment of high-energy physics experiments (say, a particle detector), detector parts will be necessarily conceptualized differently depending on the kind of information system in which they are represented. For instance, in a CAD system that is used for designing the particle detector, parts will be spatial structures; in a construction management system, they will have to be represented as tree-like structures modeling compositions of parts and their sub-parts, and in simulation and experimental data taking, parts have to be aggregated by associated sensors (readout channels), with respect to which an experiment becomes a topological structure largely distinct from the one of the design drawing. We believe that such differences also lead to different views on the knowledge level, and certainly lead to different database schemata".
- data uncertainty is reported in different ways by different research communities.
- different data formats are adopted by different research communities.
- silos in modeling data sets; and
- different concepts of what to include in the metadata.

Therefore, the traditional approach to data integration based on the design of a global schema is unfeasible, also, in the third application scenario. In the next section we will outline a new paradigm of research data integration that is well suited to the way researchers are seeking scientific information.

4 Big Research Data Integration: A New Paradigm

TIn Section 3 we have sketched some characteristics of the Research Application Scenario that heavily influence the data integration process. In this Section, we extend the description of this Scenario with some other characteristics that are equally relevant for the data integration process. These characteristics are instrumental in the re-formulation of the data integration problem. The Big Research Data era is characterized by:

– huge volumes of data available in many fields of science;
– an increasingly production of new data types that augments the complexity of data sets;
– a worldwide distribution of data sets;
– data sets with high dynamism, uncertainty, exhaustivity, and relationality.

All these characteristics of big data have contributed to the emergence of new paradigms of seeking data and creating knowledge. We have, already, described in the previous Section the new paradigm of data seeking, that is, the data exploration. A new empiricist epistemological method [15] for creating new knowledge is also emerging. In the traditional scientific method, i.e., hypothesis driven research, the data are analyzed with a specific question in mind, that is, a hypothesis. In essence, this scientific method adopts a deductive reasoning for discovering new insights from the data. In the new empiricist method, i.e., data driven research, the data are analysed with no specific question in mind. In essence, huge volumes of data together with powerful analytic tools enable data to speak for themselves. Mining big data can reveal relationships and correlations that researchers did not even know to look for. In this method an inductive or abductive reasoning is adopted for discovering new insights from the data. These two new paradigms, i.e., data exploration and data driven research, heavily influence the data integration process too. The exploratory approach to data seeking suggests the possibility for researchers to start browsing in one data set and then navigating along links into related data sets, or to support data search engines to crawl the data space by following links between data sets The empiricist method entails the integration logic that must guide the creation of these links. In fact, in the hypothesis driven method a link between two data sets is established only when a semantic relationship between variables within these data sets, dictated by the hypothesis, holds. In the empiricist method a link between two data sets is established only when a correlation between variables within these data sets, is found. In essence, in the traditional approach the integration logic allows researchers to test a theory by analysing relevant data linked together on the basis of a deductive reasoning. In the empiricist approach, the integration logic enables researchers to

discover new insights by analyzing data linked together on the basis of a inductive/abductive reasoning. Different kinds of logic (conventional logic, modal logic, causal logic, temporal logic, etc.) can be explored. An additional consideration that has an impact on the relationality/connectivity of a data set regards the properties of the data set relationships. They can be direct or indirect. A direct relationship between variables of two data sets can be recognized if these variables represent closely related concepts on the basis of the adopted logic reasoning. An indirect relationship between variables of two data sets can be established if they are directly related to a third mediating data set. Indirect relationships are based on direct relationships that enjoy the transitivity property, for example, the "causality" relationship is transitive. The indirect relationships increase the relationality/connectivity of the data sets and can enhance the data integration process and consequently the knowledge creation process.

The above two new paradigms make feasible the realization of significant advances in many scientific disciplines in the big data era as such advances can be driven by patterns in a data space. In fact, insights are arising from connections and correlations found between diverse types of data sets acquired from various modalities. Discovering semantic relationships between data sets enables new knowledge creation. The role of a data integration system is to make explicit hidden semantic relationships between data sets. Several types of semantic relationships can exist between object descriptions represented in database views. Examples of semantic relationships include: the inclusion relationship that is the standard subtype/supertype relationship; is-a and part-of relationships; member-collection relationship (association relationship); feature-event relationship; phase-activity relationship; place-area relationship; component-object relationship; antonyms/synonyms relationships, etc. Other types of semantic relationships can exist that are domain-specific. Making explicit hidden semantic relationships implies the creation of links between data sets. This process entails the creation of linked data spaces. Such linked data spaces can be implemented by exploiting linking technologies that allow to connect/link semantically related data sets. For example, data sets produced worldwide and related to the same phenomenon could be linked together creating, thus, linked data spaces in the form of thematic graphs. Researchers, interested in discovering correlations and semantic relationships between data sets contained in linked data spaces, should go through these thematic graphs. In essence, now the researchers have to explore a linked research data space by navigating through it. Based on all the above considerations the data integration paradigm problem can be re-formulated as follows:

Given a number of distributed heterogeneous and time varying data sets, link them on the basis of existing semantic and temporal relationships among them.

5 Enabling technologies

IIn this section we briefly describe the main technologies that enable the implementation of the new paradigm of big research data integration as reformulated above. These technologies (Linked Open Data, Semantic Web technologies, vocabularies etc.) are tools which may be adequate to the realization of the paradigm or in some case may need to be extended the paradigm.

5.1 Data Abstraction/ Database View

As, already, said research databases contain huge amounts of data. Usually, researchers are interested only in some parts of a database. These parts of a database (called sub datasets) are known as database views. A database view can be defined as a function [4] that, when applied to an instance (database) of a given database schema, produces a database in some other schema. In addition, the input and output schemata of this function could be represented in different data models. We think that each large database should be endowed with a

number of (possibly overlapping) views. Therefore, linking database views, distributed worldwide, on the basis of semantic relationships existing between them will be instrumental in knowledge creation.

5.2 Data Citation

Citation systems [21] are of paramount importance for the discovery of knowledge in science as well as for the reproducibility of research outcomes. Indeed, being able to cite a research data set enables potential users to discover, access, understand and reproduce it. A citation is a collection of "snippets" of information (such as authorship, title, ownership, and date) that are specified by the administrator of the data set and that may be prescribed by some standards. In essence, a "snippet" of information constitutes the metadata that must be associated with each cited data set. A data citation capability should guarantee the uniquely identification of a data set. The unique identification is achieved by using persistent identifiers (PID) such as the Life Science Identifiers (LSID), the Digital Object Identifier (DOI), the Uniform Resource Name(URN), etc. A data citation capability should also guarantee that a data citation remains consistent over time, i.e., it has to show way to the original cited data set. Guaranteeing the persistence of a data citation is demanding when the data set to be cited evolves over time [20].

5.3 Semantic Web Technologies

Linked Open Data. The term Linked Open Data refers to a set of best practices for publishing structured data on the Web [13]. In particular, Linked Data provides (i) a unifying data model. Linked Data relies on Resource Description Framework RDF as a single, unifying model; (ii) a standardized data access mechanism. Linked Data commits itself to a specific pattern of using the HTTP protocol; (iii) hyperlink-based data discovery. By using URIs as global identifiers for entities, Linked Data allow hyperlinks to be set between entities in different data sources; and (iv) self-descriptive data. A grassroots effort, the Linked Open Data, is aiming to publish and interlink open license data sets from different data sources as Linked Data on the Web.

Resource Description Framework. Resource Description Framework (RDF) [18] is a language for representing information about resources that can be identified on the Web. It represents information as node-and-arc-labeled directed graphs. The data model is designed for the integrated representation of information that originates from multiple sources, is heterogeneously structured, and is represented using different schemata. RDF aims at being employed as a lingua franca, capable of moderating between other data models that are used on the Web. In RDF, a description of a resource is represented as a number of triples. The three parts of each triple are called its subject, predicate, and object.

Time in RDF. As most of the data in the Web are time varying, there is a need for representing temporal information in RDF. This will enable users to navigate in RDF graphs across time. Therefore, it will support queries that ask for past states of the data represented by the RDF graphs. Two main mechanisms for adding temporal information in RDF graphs have been proposed in the literature [11]. The first mechanism consists in time-stamping the RDF triples that are destined to change, i.e., adding a temporal element t that labels an RDF triple. The second mechanism consists in creating a new version of the RDF graph each time an RDF triple is changed.

Named RDF Graphs. The Named Graphs data model has been introduced in order to allow a more efficient representation of metadata information about RDF data and a globally unique identification of RDF data. It is a simple variation of the RDF data model. The main idea of the model is the introduction of a naming mechanism that allows RDF triples to talk about RDF graphs. A named graph is an entity that consists of an RDF graph and a name in the form of an URI reference.

6 A Generic Scientific Application Scenario

In this section a generic scientific application scenario is described and the data integration problem is refined. In this hypothetical application environment, there are n databases distributed worldwide containing heterogeneous data represented in different formats and managed by different data management systems. These data are produced by different research teams, each following its own practices and protocols:

DB1 DB2, …, DBn

Let's, also, suppose that several database views are defined on top of each database.

We further suppose that, by using special SW, a virtual layer consisting of RDF graphs is produced on top of these views. For example, we can produce an RDF view of a relational database schema [2]. A database view has an intention and an extension. The intention describes the semantics of the view, i.e., a data query embodying the schema of the database that is created when the query is applied to a database. The extension is the data subset defined by the schema of the view, that is, the data subset selected by the query each time the query is applied. For each view, there can be any number of extensions, each produced by applying the query at a different time. Each extension has the schema defined by the intension of the view. Views should be endowed with an identifier and domain-specific metadata. We suppose that the extensional definition of the database views varies over time while the intentional definition remains unchanged. Suppose that on top of each database a number of views Views (i) are defined. The notation View1(ti) indicates that the view named View1 was

created at time ti. Therefore, the views

View1(ti), View1(ti+x), View1(ti+x+1), …, View1(ti+x+n)

have all the same schema. The dynamic nature of the extensions of database views requires mechanisms that allow the tracing of all changes that occurred during the data subset life cycle. Such mechanisms should allow, when a database operation (insert/update/delete) modifies a database subset at a given time, the time stamping of this operation and the creation of a new version of the affected extension. In essence, these mechanisms should allow the creation of findable versions of a data subset. In order to implement such mechanisms, all these time marked database operations should be kept in a persistent store in order to maintain their history with the original data subset values and an appropriate approach to data versioning should be adopted. Several approaches have been proposed in literature for the implementation of these mechanisms. In summary, upon issuing a database operation, at time ti, that modifies a database subset the following actions should be carried out: (i) the database operation is time stamped with ti; (ii) this operation is inserted and maintained in a persistent store; (iii) a new version of the affected data subset is created, i.e., a new extension of each view defined on the updated data set; (iv) each newly generated extension is also time-stamped with ti. Therefore, a view is associated with different data operations that have affected its extensions. Both view and operations are endowed with time stamps that indicate the time of the execution of the data operation, i.e., the time of the modification of the view extension. The natural ordering of the time stamps associated with the operations that affect the extension of a view induces a natural ordering of the extensions produced by these operations, so that an extension can be seen as a version of the extension that precedes it in that ordering. Therefore, the time stamped views constitute a directed acyclic graph (DAG) whose nodes are the views at different points of time [View1(ti), View1(ti+x),…..,View1(ti+x+n)] and the links represent temporal relationships between these nodes. A DAG has a topological ordering, a sequence of the nodes such that every arc is directed from earlier to later in the sequence. In essence, a DAG represents the evolution of a data subset over time. Finding a particular extension of a database view, i.e., the version of its extension at time ti implies (i) to identify the view through its ID; (ii) to cross the acyclic graph of this view until reaching the view View1(ti); and (iii) to execute the database operation associated with the View1(ti). The data integration problem can be slightly refined with respect to the formulation given in Section 4 as follows:

Given a number of distributed heterogeneous and time varying database views, represented by directed acyclic graphs, link them on the basis of existing semantic and temporal relationships among them.

To implement this new paradigm of data integration, we propose to proceed according to the following two steps: The first step consists in creating on top of each database view schema a level of virtual description. This virtualization

describes the intentional part of a data subset (its schema) in terms of the RDF data model. Therefore, all database view schemata are represented in the RDF model while their extensions, i.e., the data subsets defined by them remain expressed in the data models supported by the local data management systems. This level of virtualization permits the uniform representation of the different schemata of the database views. Therefore, it allows researchers to query the database views and access their extensions (data subsets) by using a unique query language like SPARQL. This frees the researchers from the need to know the different query languages supported by the local database management systems in order to query the single database views.

The second step consists in inserting the virtualized database views together with their IDs and metadata in domain-specific registries.

The third step consists in linking the several virtualized database view schemata, described in these registries, on the basis of existing semantic and temporal relationships between them. The linking operation is guided by the adopted integration logic and produces a linked database view space where thematic graphs can be designed. These graphs constitute patterns of data that transform data in knowledge.

Finally, researchers by using appropriate query languages, as for example SPARQL, can explore the linked space by following appropriate links.

7 Conclusions

This paper presents a vision and outlines a direction to be followed in order to solve the data integration problem in the new emerging scientific data context. We identified the challenges that must be faced in order to successfully implement the new big research data integration paradigm that can be formulated as follows: "Given a number of distributed heterogeneous and time varying data sets, link them on the basis of existing semantic and temporal relationships among them." The main challenges we identified are:

- making the schema of a database view citable;
- endowing the schema of a database view with an identifier, a time stamp and appropriate metadata;
- creating mechanisms that support the versioning of the data subsets defined by the schemata of database views;
- developing mappings that allow an RDF application to access data subsets of non-RDF database views without having the need of transforming the data subsets into RDF triples;
- introducing temporal information into RDF graphs in order to be able to query this type of information;
- adopting an appropriate integration logic that has to guide a search engine in the identification of semantic and/or temporal relationships between distributed worldwide database view schemata and creating f links among them;

– developing or using existing domain-specific vocabularies/ontologies to support the process of semantically and temporarily linkage of database views;
– developing or using existing Catalogues/Registries where database view schemata are published;
– developing of query languages or extension of existing ones that allow to traverse linked RDF graphs.

References

1. Bernstein, P.A., Haas, L.M.: Information integration in the enterprise. Communications of the ACM **51**(9), 72–79 (2008)
2. Bizer, C., Seaborne, A.: D2rq-treating non-rdf databases as virtual rdf graphs. In: Proceedings of the 3rd international semantic web conference (ISWC2004). vol. 2004. Proceedings of ISWC2004 (2004)
3. Brackett, M.H.: Data Resource Integration: Understanding and Resolving a Disparate Data Resource, vol. 2. Technics Publications (2012)
4. Buneman, P., Davidson, S., Frew, J.: Why data citation is a computational problem. Communications of the ACM **59**(9), 50–57 (2016)
5. Chawathe, S., Garcia-Molina, H., Hammer, J., Ireland, K., Papakonstantinou, Y., Ullman, J., Widom, J.: The tsimmis project: Integration of heterogenous information sources (1994)
6. Council, N.R., et al.: Steps toward large-scale data integration in the sciences: Summary of a workshop. National Academies Press (2010)
7. Daraio, C., Lenzerini, M., Leporelli, C., Moed, H.F., Naggar, P., Bonaccorsi, A., Bartolucci, A.: Data integration for research and innovation policy: an ontology-based data management approach. Scientometrics **106**(2), 857–871 (2016)
8. Doan, A., Halevy, A.Y.: Semantic integration research in the database community: A brief survey. AI magazine **26**(1), 83–83 (2005)
9. Dong, X.L., Srivastava, D.: Big data integration. In: 2013 IEEE 29th international conference on data engineering (ICDE). pp. 1245–1248. IEEE (2013)
10. Guarino, N., Oberle, D., Staab, S.: What is an ontology? In: Handbook on ontologies, pp. 1–17. Springer (2009)
11. Gutierrez, C., Hurtado, C.A., Vaisman, A.: Introducing time into rdf. IEEE Transactions on Knowledge and Data Engineering **19**(2) (2007)
12. Halevy, A., Rajaraman, A., Ordille, J.: Data integration: the teenage years. In: Proceedings of the 32nd international conference on Very large data bases. pp. 9–16. VLDB Endowment (2006)
13. Heath, T., Bizer, C.: Linked data: Evolving the web into a global data space. Synthesis lectures on the semantic web: theory and technology **1**(1), 1–136 (2011)
14. Idreos, S., Papaemmanouil, O., Chaudhuri, S.: Overview of data exploration techniques. In: Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data. pp. 277–281. ACM (2015)
15. Kitchin, R.: Big data, new epistemologies and paradigm shifts. Big Data & Society **1**(1), 2053951714528481 (2014)
16. Koch, C.: Data integration against multiple evolving autonomous schemata. Ph.D. thesis, Vienna U. (2001)
17. Lenzerini, M.: 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 (2002)

18. McBride, B.: The resource description framework (rdf) and its vocabulary description language rdfs. In: Handbook on ontologies, pp. 51–65. Springer (2004)
19. Naumann, F., Bilke, A., Bleiholder, J., Weis, M.: Data fusion in three steps: Resolving inconsistencies at schema-, tuple-, and value-level. IEEE Data Engineering Bulletin **29**(2), 21–31 (2006)
20. Proll, S., Rauber, A.: Scalable data citation in dynamic, large databases: Model and reference implementation. In: Big Data, 2013 IEEE International Conference on. pp. 307–312. IEEE (2013)
21. Silvello, G.: Theory and practice of data citation. Journal of the Association for Information Science and Technology **69**(1), 6–20 (2018)
22. Vassiliadis, P.: A survey of extract–transform–load technology. International Journal of Data Warehousing and Mining (IJDWM) **5**(3), 1–27 (2009)
23. Wiederhold, G.: Interoperation, mediation and ontologies. In: FGCS Workshop on Heterogeneous Cooperative Knowledge-Bases (1994)
24. Ziegler, P., Dittrich, K.R.: Three decades of data intecration—all problems solved? In: Building the Information Society, pp. 3–12. Springer (2004)