Data Integration: state of the art, new issues and research plan

Marco Ruzzi
Dipartimento di Informatica e Sistemistica
`ruzzi@dis.uniroma1.it`

October 21, 2004

Abstract

In this report we present a summary of the recent literature in the Data Integration research area, focusing our attention on some new issues that are relevant to this area, mostly from an architectural perspective. In fact, although general architectures for data integration systems assign them a centralized role, it does not seem to be a valid approach in real distributed scenarios: some recent works propose new architectures for data integration, deployed in a *peer-to-peer* environment. We will also summarize some results produced in the first year of the Ph.D. course and we will set out a research plan for the rest of such course.

1 Introduction

Today's fast and continuous growth of large business organizations, often deriving from mergers of smaller enterprises, enforces an increasing need in integrating and sharing large amounts of data, coming from a number of heterogeneous and distributed data sources. *Data integration systems* aim to fulfill such need by suitably querying the sources in their own language, retrieving the relevant data and reassembling them into the answers to the user queries. All these issues are concerned with the Data Integration problem.

One of the most natural way to think about a data integration system, consists of a uniform mediation layer placed between the user and the data sources. The user interacts with the system by querying a *global schema* [27], i.e., a virtual representation of the data stored at the sources, and the system will carry out the task of dealing with the sources to retrieve the information satisfying the user request. On the other hand, the data integration system will keep an internal representation of the data at the sources, called *source schema* [27], so as the relationships subsisting between entities of the global and the source schema: such relationships are represented by the *mapping* [27]. The interaction between the data integration system and the sources is made easy by some software modules, called *wrappers*, that hide the native characteristics of the sources, allowing the system to converse with it using its language.

To better represent the users' interest domain, integrity constraints can be expressed over the global schema [6]. The system will consider such constraints when answering the user queries, providing answers that have to be coherent with respect to such constraints.

Some recent developments of data integration research, enforce an inadequacy of such a centralized architecture in satisfying many of the real integration cases. Indeed, as we will explain better in the following, the design and maintenance of the global schema and the mapping constitutes a very hard task, also considering mutable settings in which sources can join and leave the system at any time. Furthermore, stressing the abstraction level, one can think of an environment in which sources may act as provider of information, as well as consumer.

This short report is structured as follows: in Section 2 we present the basic concepts regarding data integration referring to the centralized architecture, pointing out some limitations of such an approach; in Section 2.1 we show several recent results about the semantic data integration; in Section 3 we introduce some new architectural approaches that have been presented in the recent literature, while in Section 4 we present the results of the first year of Ph.D. course and some future plan.

2 Data Integration: classical framework

In this section we summarize the main results of the data integration community referring to a classical architecture, constituted by a centralized virtual global schema that provides a mediated representation of the sources integrated by the system. To better expose such results, we adopt the relational model [16] to represent global concepts as well as source entities; moreover, we represent the mapping between source and global relations by means of conjunctive queries [1].

Formally, a *data integration system* $\mathcal{I}$ [27], is a triple $< \mathcal{G}, \mathcal{S}, \mathcal{M} >$ in which:

- $\mathcal{G}$ is the *global schema*, i.e., a set of global relational symbols, each one with an associated arity (the number of its attributes), plus a set of integrity constraints expressed over such relational symbols;
- $\mathcal{S}$ is the *source schema*, i.e., a set of relational symbols (disjoint from $\mathcal{G}$), that constitutes a relational representation of the data stored at the sources;
- $\mathcal{M}$ is the *mapping* between $\mathcal{G}$ and $\mathcal{S}$, constituted by a set of assertions of the form $\{q_{\mathcal{S}},q_{\mathcal{G}}\}$, in which $q_{\mathcal{S}}$ is a conjunctive query over the global schema, while $q_{\mathcal{G}}$ is a conjunctive query over the global schema.

The semantics of the system is given in terms of the data at the sources: starting from a database instance $\mathcal{D}$ for the source schema, we define the semantics of a data integration system $\mathcal{I}$, as the set of all the global database $\mathcal{B}$ that respect the following conditions:

- $\mathcal{B}$ is a legal database instance for $\mathcal{G}$, that is, respects the relational structure of $\mathcal{G}$, and satisfies the integrity constraints expressed on it;
- $\mathcal{B}$ satisfies the mapping assertions with respect to the source instance $\mathcal{D}$. We remand to [27] for any further details.

With respect to the mapping assertions, different assumptions can be done, that affect the notion of satisfaction of mapping. In particular, if we assume that the mapping is *sound*, then we have that the data provided by the sources are a subset of the global data – the extension of $q_{\mathcal{S}}$ is contained into the extension of $q_{\mathcal{G}}$. Conversely, if the mapping is considered to be *complete*, the data provided by the sources are a superset of the global data – the extension of $q_{\mathcal{S}}$ contains the extension of $q_{\mathcal{G}}$. Finally we assume that a mapping is *exact*, when it is both sound and complete. We point out that, due to the general characteristics of the sources, that are distributed, autonomous and independent, the sound mapping assumption is more reasonable in a data integration environments [27].

The mapping design is one of the crucial tasks in defining a data integration system specification. In fact, tuning the shape of the mapping queries, one can obtain different representations that have well-known proprieties. We say that a mapping assertion follows the *global-as-view (GAV)* paradigm, when $q_{\mathcal{G}}$ corresponds to a full query over a single global relation: an assertion of that kind gives a straightforward specification of the global data, in terms of the source data. Dually, the *local-as-view (LAV)* approach, let us define $q_{\mathcal{S}}$ as a full query over a single source relation. Both formalisms present advantages and drawbacks [31]. GAV mappings ease the query answering process that can be done by means of simple unfolding techniques, but its structure is not well suited for adding or removing sources: every change in the source schema may lead to mapping redesign. LAV mappings are better from this perspective, because adding or removing a source specification, only involves adding or removing a single mapping assertion. On the other hand, query answering in LAV data integration systems is an hard task, because the mapping views only provide a partial information about the elements of the global schema. In [29] the authors present the MiniCon algorithm that efficiently solves the problem of query answering in LAV data integration systems. [31] presents and compares TSIMMIS, a GAV data integration system, and Information Manifold, a LAV one. Other examples of GAV information systems are Garlic [15], MOMIS [2], IBIS [7] and DIS@DIS that we will present in Section 4. As far as we know IBIS and DIS@DIS are the only systems that are able to deal with integrity constraints.

As we will see better in the Section 3, there is a more general approach in mapping design, that generalizes both the GAV and the LAV paradigms. Such an approach, called *generalized local-as-view (GLAV)*, requires the designer to associate a general query over the source relations $q_{\mathcal{S}}$ to a general query $q_{\mathcal{G}}$ over the global relations. Section 3 will point out that GLAV mappings are more expressive, and are well suited to represent complex relationships in distributed data integration environments.

2.1 Recent approaches for semantic Data Integration

One of the main challenges in Data Integration is to deal with integrity constraints expressed over the global schema. In fact, when a system integrates data coming from different independent information sources, reconciliated data may not satisfy the integrity constraints imposed over the global schema. More, many of the methodologies to deal with these problems are employed in the reconciliation step with ad-hoc data cleaning algorithms.

Under the assumption of sound mapping, the repair of a global inconsistent instance can be produced by only adding tuples. For example, if an inclusion dependencies is violated, it is possible to build a repair by adding tuples to the inconsistent instance [6]. However, this assumption presents some serious drawbacks: the repair may be infinite, making the query answering undecidable, and does not allow for the deletion of tuples (needed to repair key constraint violation). Some recent works [9, 23, 5, 22] show that in many cases, the sound assumption on the mapping is too binding, and may lead to situations in which the semantics of the data integration system is empty, i.e., there are not legal global instances, making the query answering meaningless. The authors propose a new semantic approach whose main characteristic is to consider repairs that satisfy the integrity constraints over the global schema and that approximate at best the satisfaction of the sound assumption. This semantics, called *loosely-sound semantics*, constitutes a less strict assumption on data and allows one to add as well as delete tuples to build the repair. Even though the semantics proposed concerns databases with inconsistent information, the technique can also be applied to a data integration environment [10]. The authors also show that query answering under the loosely-sound semantics is coNP-complete with respect to data complexity in the presence of keys and foreign keys over the global schema.

As presented in [21], another way to define a semantic for inconsistency, is to state some preference assertions between data sources. The basic idea, mostly useful in a data integration environment, is that best sources should be preferred to the others. In [21] the authors define the *maximally-sound semantics* that takes into account a set of source preference expression when defining the global interpretation of the source data.

Before concluding this section, we point out that several authors are investigating the problem of finding out particular classes of queries for which query answering over inconsistent databases can be done in polynomial time, also in the presence of key and foreign key [20].

3 New architectures for Data Integration

Some recent developments of infrastructures and algorithms for distributed systems, and also the fast growth of the number of heterogeneous information sources scattered on the web, have steadily affected the research course of the last years. It derives an inadequacy of the centralized architecture introduced so far to satisfy real integration requests. Moreover, in a continuously changing scenario in which sources may join and leave at any time, the design and maintenance of a global schema and a mapping are too expensive. In this section we present a new direction of the data integration science, oriented towards an architectural extension of its well known techniques.

3.1 Peer-to-peer Data Management Systems (PDMS)

The peer-to-peer (from now on P2P) paradigm consists of a wide network of interconnected computational *peers*, or information nodes, that cooperate with each other exchanging services and information [3]. Each peer shares a part or all of its resources with the network and also benefit of resources offered by other peers. The success of the P2P architecture is determined by several advantages: scalability, robustness and lack of administration needs are some of the main properties that a P2P system possesses. In general the wealth of such systems grows with the number of participants, that increase the memory for data storage, and the computational power of the whole system.

However, as pointed out in [24], often generic P2P systems do not take care of the semantics of the data exchanged. This is a serious drawback, mostly considering that when the network grows, it becomes hard to predict the location and the quality of the data provided by the system. Therefore, there are several problems about availability and consistency of the services provided by the P2P system. In the same paper, the authors identify the *data placement problem*, that is, how to distribute data and work so as to the full query workload is covered with lower cost under the existing resource constraints.

A *P2P data management system* aims to solve these issues. The main goal of such systems is to provide semantic interoperability in the absence of a global information schema. All the knowledge about the topology and extension of the network resides at the peers. In fact, every peer is interconnected with other peers by means of *coordination formulas* [3], that allow every single peer to exploit acquaintances coming from other peers. A P2P system decomposes the user queries by recursively applying coordination formulas, that may also act as a kind of constraints for the propagation of updates among the network. In [30] the authors present the *Local Relational Model (LRM)*, a data model specifically designed for P2P data management systems. Each peer has a *local database schema* whose semantics is defined on a local domain (disjoint from the domain of other peers). The information propagates between peers thanks to coordination rules and to relations between domains of different peers. Such relations express overlapping between local databases of different peers, i.e., different constants that represent the same object. In such a scenario, coordination formulas are used, in a declarative fashion, to state semantic inter-dependencies between two different local databases. One of the main peculiarities of such formalism is that there is not a concept of global consistency, but only local consistency at peer level.

Many authors have deeply investigated the issues arising in P2P distributed environments [26, 19]. An interesting topic is how to express the logical interconnections between peers. Many of the sources available on the web, for example, cannot be modelled as relational databases, but rather as *data webs* [19] with a set of *entry points*. Under this perspective, the resource scenario can be seen as a set of data webs that can be modelled together as a *web schema*, i.e., a direct graph in which data webs are nodes, and links between them are arcs. In [19] the authors provide a formalization of web schemas architecture, providing also techniques for query answering: a data web can be queried if it has an entry point, and links between data webs can be used to "navigate" through the web schema. Starting from the user query, the technique proposed produces a *navigational plan*, which is then translated into an extension of relational algebra through the *traverse* operator, that allows to traverse links across different data webs. The authors also point out that the LAV and GAV formalism are inadequate to suitably represent link between nodes: in fact they adopt the GLAV paradigm, introduced in Section 2, that generalizes both the approaches. In that paper it is shown that GLAV mapping reaches the tradeoff limit between expressive power and tractability of query answering. It derives that in these settings the complexity of query answering is polynomial in data complexity.

To conclude this short survey, we underline that in many real integration cases, the adoption of database driven techniques can hinder some lightweight data storage task: these issues are addressed in [26]. The scenario proposed there contemplates a peer schema in which peers hold a set of *peer relations*, to represent its data, together with a set of *stored relations*, analogous to the sources of classical data integration systems. Moreover, every peer holds two kinds of peer mappings to interlink its schema with the schema of other peers: *inclusion and equality mappings*, similar to LAV classical mapping, and *definitional mapping*, analogous to GAV ones. The semantic of the whole PDMS is given based on a data instance, that is an extension of the relation of each peer. The consistency of such an instance depends on the satisfaction of inclusion and equality mappings. The authors show that the query answering problem under such assumptions is in general undecidable, due to cycles among definitional mappings. However, bounding the topology of cycles, a polynomially tractable algorithm can be defined, that is able to chain between different peers mapping to get the answer to the query.

3.2 Peer-to-peer Data Integration

All the approaches presented in the previous section are not able to deal with an arbitrary topology of peer interconnections, due to the particular semantics adopted in defining query answering techniques. Nevertheless, imposing limitations on peer interconnections does not seem to be a valid approach. The results presented in [13] solve that problem. The authors underline that assigning a global semantic to a distributed environment, leads to undecidability of query answering. This is shown in very simple settings, in which only three peers operate. The solution proposed consists in the adoption of an epistemic semantics instead of the classical first order logic semantic. Under this new semantics, mapping assertions, expressed in GLAV, are interpreted so that only the knowledge of the peer is transferred to other peers. In the same work, the authors define a distributed algorithm for query answering that exploits a transaction mechanism to ensure the semantic correctness of the whole process with respect to the epistemic semantics.

The same authors have extended the work providing an infrastructure of P2P data integration techniques for virtual organizations, implemented on Data Grids [14]. The deployment proposed there model a P2P system as a set of *data peers* and *hyper peers*. The formers are systems that export data in terms of an exported schema, while the latter do not export data, but are interconnected both with data and hyper peers, building up the topology of the network. An hyper peer connected to other data peers correspond to a classical GLAV data integration system, and its semantics is defined resorting to epistemic logic [13]. Query answering in the Hyper framework is performed by splitting each GLAV mapping assertion into two halves: a LAV assertion and a GAV assertion, linked by means of a new relational symbol. Such LAV assertions are used to produce a logic program exploiting algorithms for query answering using views, while the GAV assertions are used by the system to generate the extensions on which the logic program can be evaluated.

3.3 Other approaches to Data Integration

To conclude this section, we mention some other approaches to the problem of Data Integration. The *what-to-ask (WTA)* problem, introduced in [12] is the problem of answering the queries posed to a peer by only relying on the query answering service of the peers. In particular, the setting proposed in [12] contemplates only two peers: the local and the remote peer. That work presents a formalization for the WTA problem in such a framework, providing also a solution for that problem when a basic ontology language is used to express the knowledge base of the two peers. A solution for the WTA problem consists in what one has to ask to the remote peer in order to answer queries posed to the local peer. More, the authors show that, when the expressive power of the ontology language is enriched, the WTA problem is not always decidable. The algorithm proposed can be seen as a new technique for query rewriting that can be exploited also for data integration systems.

The data integration approach aims to answer the user queries taking advantage of several query rewriting techniques. This means that the data are only at the sources and the system is not in charge to materialize them locally (virtual data integration). Conversely, *data exchange* [18] solutions adopt a strategy of materialization of the remote data. In data exchange we have a remote source schema, i.e., a set of information sources, and a target schema, analogous to the global schema of data integration systems, that has to be materialized. The data exchange problem consists in finding an instance of the target schema, starting from an instance of the remote source schema. Such a problem, that has in general several solutions, is addressed in [17], where the authors presents a formalization of the semantic and query answering for the problem. They define the *universal solution*, i.e., a solution that is homomorphic with all the other solutions, and hence is the most general solution. However, since there can be more than one universal solution, a minimality criterion is used to find the best one. It is possible to identify a substructure common to all universal solutions that is also isomorphic to them. This substructure, called the *core*, can be computed exploiting well-known algorithm of graph theory that, in the data exchange setting, can be executed in polynomial time in data complexity.

4 Research results and future plan

Our skill in data integration has been achieved by designing and implementing DIS@DIS [8], a system for semantic data integration under integrity constraints. The development of such system was started thanks to the work of some students during their master degree thesis that provide an implementation of some algorithms presented in [9]. In particular, DIS@DIS handles GAV data integration systems providing correct answers to queries posed over a relational global schema enriched with key and foreign key. As a central feature, DIS@DIS exploits some reasoning techniques involving inclusion dependencies and their interaction with exclusion dependencies. Moreover, the system implements some query rewriting techniques that only operate at the intensional level, ensuring good performances on real cases [11]. Many modules of the system have been exploited in the INFOMIX European research project.

Some work has also been spent finding a way to deploy data integration systems relying upon a commercial tool for data federation [25]. In particular, two techniques have been proposed [28], that aim to solve such issue by compiling the data integration system specification into IBM DB2 Information Integrator tool. Together with the compilation techniques mentioned so far, also two query translation techniques are presented, that handle user queries so as to be issued on the compiled specification. The experimental results achieved during the test of the two approaches suggest some new research ideas for this topic, with respect to the possibility of taking care of integrity and typing constraints in such settings.

All the discussion made so far points out two main research lines for data integration that we consider as interesting. First, the need in formal and efficient techniques for semantic data integration. Second, it seems to be natural to extend semantic data integration approaches to the case of P2P architectures. In fact, as far as we know, only in [4] a technique is proposed to face the query answering problem in a distributed data exchange environment, when integrity constraints are expressed over the peer definitions.

References

[1] Serge Abiteboul, Richard Hull, and Victor Vianu. *Foundations of Databases.* Addison Wesley Publ. Co., Reading, Massachussetts, 1995.

[2] D. Beneventano, S. Bergamaschi, S. Castano, A. Corni, R. Guidetti, G. Malvezzi, M. Melchiori, and M. Vincini. Information integration: the MOMIS project demonstration. In *Proc. of the 26th Int. Conf. on Very Large Data Bases (VLDB 2000)*, 2000.

[3] Philip A. Bernstein, Fausto Giunchiglia, Anastasios Kementsietsidis, John Mylopoulos, Luciano Serafini, and Ilya Zaihrayeu. Data management for peer-to-peer computing: A vision. In *Proc. of the 5th Int. Workshop on the Web and Databases (WebDB 2002)*, 2002.

[4] Leopoldo Bertossi and Loreto BRavo. Query answering in data esxchange peer-to-peer systems. In *Proc. of the 10th Int. Conf. on Database Theory (ICDT 2004)*, 2004.

[5] Loreto Bravo and Leopoldo Bertossi. Logic programming for consistently querying data integration systems. In *Proc. of the 18th Int. Joint Conf. on Artificial Intelligence (IJCAI 2003)*, pages 10–15, 2003.

[6] Andrea Calì, Diego Calvanese, Giuseppe De Giacomo, and Maurizio Lenzerini. Data integration under integrity constraints. In *Proc. of the 14th Int. Conf. on Advanced Information Systems Engineering (CAiSE 2002)*, volume 2348 of *Lecture Notes in Computer Science*, pages 262–279. Springer, 2002.

[7] Andrea Calì, Diego Calvanese, Giuseppe De Giacomo, Maurizio Lenzerini, Paolo Naggar, and Fabio Vernacotola. IBIS: Data integration at work. In *Proc. of the 10th Ital. Conf. on Database Systems (SEBD 2002)*, pages 291–298, 2002.

[8] Andrea Calì, Saverio De Nigris, Domenico Lembo, Gabriele Messineo, Riccardo Rosati, and Marco Ruzzi. DIS@DIS: a system for semantic

data integration under integrity constraints. In *Proc. of the 4th Int. Conf. on Web Information Systems Engineering (WISE 2003)*, pages 335–338, 2003.

[9] Andrea Calì, Domenico Lembo, and Riccardo Rosati. On the decidability and complexity of query answering over inconsistent and incomplete databases. In *Proc. of the 22nd ACM SIGACT SIGMOD SIGART Symp. on Principles of Database Systems (PODS 2003)*, pages 260–271, 2003.

[10] Andrea Calì, Domenico Lembo, and Riccardo Rosati. Query rewriting and answering under constraints in data integration systems. In *Proc. of the 18th Int. Joint Conf. on Artificial Intelligence (IJCAI 2003)*, pages 16–21, 2003.

[11] Andrea Calì, Domenico Lembo, Riccardo Rosati, and Marco Ruzzi. Experimenting data integration with DIS@DIS. In *Conference on Advanced Information Systems Engineering*, 2004. To appear.

[12] Diego Calvanese, Giuseppe De Giacomo, Domenico Lembo, Maurizio Lenzerini, and Riccardo Rosati. What to ask to a peer: Ontology-based query reformulation. In *Proc. of the 9th Int. Conf. on Principles of Knowledge Representation and Reasoning (KR 2004)*, 2004.

[13] Diego Calvanese, De Giacomo Giuseppe, Maurizio Lenzerini, and Riccardo Rosati. Logical foundations of peer-to-peer data integration. In *Proc. of the 23rd ACM SIGACT SIGMOD SIGART Symp. on Principles of Database Systems (PODS 2004)*, 2004.

[14] Diego Calvanese, De Giacomo Giuseppe, Maurizio Lenzerini, Riccardo Rosati, and Guido Vetere. Hyper: a framework for peer-to-peer data integration on grids. In *Proc. of the 23rd ACM SIGACT SIGMOD SIGART Symp. on Principles of Database Systems (PODS 2004)*, 2004.

[15] M. J. Carey, L. M. Haas, P. M. Schwarz, M. Arya, W. F. Cody, R. Fagin, M. Flickner, A. Luniewski, W. Niblack, D. Petkovic, J. Thomas, J. H. Williams, and E. L. Wimmers. Towards heterogeneous multimedia information systems: The Garlic approach. In *Proc. of the 5th Int. Workshop on Research Issues in Data Engineering – Distributed Object Management (RIDE-DOM'95)*, pages 124–131. IEEE Computer Society Press, 1995.

[16] E. F. Codd. A relational model of data for large shared data banks. *Communications of the ACM*, 13(6):377–387, 1970.

[17] Ronald Fagin, Phokion G. Kolaitis, Renée J. Miller, and Lucian Popa. Data exchange: Semantics and query answering. In *Proc. of the 9th Int. Conf. on Database Theory (ICDT 2003)*, pages 207–224, 2003.

[18] Ronald Fagin, Phokion G. Kolaitis, and Lucian Popa. Data exchange: Getting to the core. In *Proc. of the 22nd ACM SIGACT SIGMOD SIGART Symp. on Principles of Database Systems (PODS 2003)*, pages 90–101, 2003.

[19] Marc Friedman, Alon Levy, and Todd Millstein. Navigational plans for data integration. In *Proc. of the 16th Nat. Conf. on Artificial Intelligence (AAAI'99)*, pages 67–73. AAAI Press/The MIT Press, 1999.

[20] Ariel D. Fuxman and Renée J. Miller. First order query rewriting for inconsistent databases. In *Proc. of the 11th Int. Conf. on Database Theory (ICDT 2005)*, 2005. To appear.

[21] Giuseppe De Giacomo, Domenico Lembo, Maurizio Lenzerini, and Riccardo Rosati. Tackling inconsistencies in data integration through source preferences. In *IQIS 2004*, 2004.

[22] Gianluigi Greco, Sergio Greco, and Ester Zumpano. A logic programming approach to the integration, repairing and querying of inconsistent databases. In *Proc. of the 17th Int. Conf. on Logic Programming (ICLP'01)*, volume 2237 of *Lecture Notes in Artificial Intelligence*, pages 348–364. Springer, 2001.

[23] Gianluigi Greco, Sergio Greco, and Ester Zumpano. A logical framework for querying and repairing inconsistent databases. *IEEE Trans. on Knowledge and Data Engineering*, 15(6):1389–1408, 2003.

[24] Steven Gribble, Alon Halevy, Zachary Ives, Maya Rodrig, and Dan Suciu. What can databases do for peer-to-peer? In *Proc. of the 4th Int. Workshop on the Web and Databases (WebDB 2001)*, 2001.

[25] L.M. Haas, E.T. Lin, and M.A. Roth. Data integration through database federation. *IBM Systems Journal*, 41(4):578–596, 2002.

[26] Alon Halevy, Zachary Ives, Dan Suciu, and Igor Tatarinov. Schema mediation in peer data management systems. In *Proc. of the 19th IEEE Int. Conf. on Data Engineering (ICDE 2003)*, pages 505–516, 2003.

[27] Maurizio Lenzerini. Data integration: A theoretical perspective. In *Proc. of the 21st ACM SIGACT SIGMOD SIGART Symp. on Principles of Database Systems (PODS 2002)*, pages 233–246, 2002.

[28] Antonella Poggi and Marco Ruzzi. Filling the gap between data federation and data integration. In *Sistemi Evoluti per Basi di Dati*, 2004.

[29] Rachel Pottinger and Alon Y. Levy. A scalable algorithm for answering queries using views. In *Proc. of the 26th Int. Conf. on Very Large Data Bases (VLDB 2000)*, pages 484–495, 2000.

[30] L. Serafini, F. Giunchiglia, J. Mylopoulos, and P. Bernstein. The local relational model: Model and proof theory, 2001.

[31] Jeffrey D. Ullman. Information integration using logical views. In *Proc. of the 6th Int. Conf. on Database Theory (ICDT'97)*, volume 1186 of *Lecture Notes in Computer Science*, pages 19–40. Springer, 1997.