Tackling Inconsistencies in Data Integration through Source Preferences

Giuseppe De Giacomo Domenico Lembo Maurizio Lenzerini Riccardo Rosati

Dipartimento di Informatica e Sistemistica
Università di Roma "La Sapienza"
Via Salaria 113, I-00198 Roma, Italy

{degiacomo,lembo,lenzerini,rosati}@dis.uniroma1.it

ABSTRACT

Dealing with inconsistencies is one the main challenges in data integration systems, where data stored in the local sources may violate integrity constraints specified at the global level. Recently, declarative approaches have been proposed to deal with such a problem. Existing declarative proposals do not take into account preference assertions specified between sources when trying to solve inconsistency. On the other hand, the designer of an integration system may often include in the specification preference rules indicating the quality of data sources. In this paper, we consider *Local-As-View* integration systems, and propose a method that allows one to assign formal semantics to a data integration system whose declarative specification includes information on source preferences. To the best of our knowledge, our approach is the first one to consider in a declarative way information on source quality for dealing with inconsistent data in Local-As-View integration systems.

Keywords

Data Integration, Inconsistent Data, Data Quality.

1. INTRODUCTION

Dealing with inconsistencies is one the main challenges in data integration [18]. The data integration systems we are interested in this work are characterized by an architecture based on a global schema and a set of sources. The sources contain the real data, while the global schema provides a reconciled, integrated, and virtual view of the underlying sources. A mapping relates data sources with the elements of the global schema. Classical approaches to specifying the mapping are the *Global-As-View* (GAV) approach, in which each global element is associated with a view over the sources, and the *Local-As-View* (LAV) approach, in which, conversely, to each source element is associated a view over the global schema. Inconsistency may arise because the global schema generally contains integrity constraints, and sources may contain data that, combined with other sources, may contradict constraints. Since one of the main goals of a data integration system is to answer queries posed in terms of the global schema, and since the answer to a query is based on the data stored in the sources, it is immediate to verify that inconsistency dramatically affects the ability of the system of providing meaningful answers to queries.

Roughly speaking, there are two approaches to deal with inconsistent data in information integration. The first approach is procedural in nature, and is based on domain-specific transformation and cleaning [6] procedures applied to the data retrieved from the sources.

The second approach is declarative. Indeed, several papers present techniques for providing informative answers even in the case of a database that does not satisfy its integrity constraints (see, for example, [2, 3, 14]). Although interesting, such results are not specifically tailored to the case of different consistent data sources that are mutually inconsistent, that is the case of interest in data integration. This case is addressed in [21], where the authors propose an operator for merging databases under constraints. Such operator allows one to obtain a maximal amount of information from each database by means of a majority criterion used in case of conflict. However, also the approach described in [21] does not take explicitly into account the notion of mapping between sources and global schema as introduced in most declarative data integration settings. Intuitively, in such settings the problem of dealing with inconsistency is particularly challenging, since integrity constraints are specified at the global level, and inconsistency may arise only because of the way source data are related with global elements by means of the mapping. Only recently, some papers [5, 10, 7, 12] have tackled data inconsistency in declarative data integration settings. Such papers get rid of inconsistency by suitably "repairing" data retrieved from the sources, according to some minimality criteria. Basically, such papers extend the studies on a single inconsistent database [2, 14] to the case of data integration.

We point out that none of the above mentioned approaches takes into account preference criteria when trying to solve inconsistencies among data sources. On the other hand, we believe that, when specifying a data integration system, the designer may often include in the specification information on sources' quality (e.g., reliability, availabil-

IQIS 2004 Maison de la Chimie, Paris, France

ity, etc.). Obviously, "best" sources should be preferred to the others when solving inconsistencies. Note that the idea of taking sources' quality into account in data integration is not new. For example, in [23], aspects related to different source quality parameters are incorporated in the query planner. However, quality information is not exploited for dealing with inconsistent source data. In [26], an approach is described in which users can indicate in their queries the quality degree of the answer that they want to get. These indications are taken into account in case of conflicting data. However, no information on the quality of the data sources is considered when computing the answers to the query. In [24], "reliability degrees" of the sources can be taken into account in the mapping definition: mapping views indicate how to resolve possible inconsistency, i.e., which source has to be favored in case of conflict. Also, in [1] an inconsistency resolution methodology is described which computes answers to user queries by applying pre-defined conflict resolution policies based on the quality of data at the sources and further quality parameters provided by the user. The latter two mentioned approaches are clearly procedural. Furthermore, they consider only GAV or limited forms of mappings.

In this paper, we address the issue of dealing with inconsistencies in LAV data integration systems specified in a declarative way, and we present a method that exploits information on source preferences. We concentrate our attention on the specification of the semantics of our approach. In other words, we propose a method that allows one to assign formal semantics to a data integration system whose declarative specification includes information about source preferences. We focus our attention on LAV systems, which have been shown superior to other approaches in supporting extensibility and flexibility. To the best of our knowledge, the semantics proposed in this paper is the first semantics for LAV data integration systems that take into account information on sources' quality for dealing with inconsistent data.

The paper is organized as follows. In Section 2 we provide a means to declaratively specify a data integration system. In Section 3 we discuss the first-order logic semantics of LAV data integration systems, and we point out its drawbacks in dealing with inconsistencies. In Section 4, we define the *maximally-sound semantics* (originally presented in [17, 10] with the name of loosely-sound semantics, although restricted to the GAV setting), and we show how it allows to overcome some of the drawbacks of the FOL semantics. In Section 5, we propose a new semantics, that combines the power of the maximally-sound semantics with information on sources' quality. Finally, Section 6 concludes the paper.

2. FRAMEWORK

In this section we define a general formal framework for data integration. Informally, a data integration system consists of a (virtual) global schema, which specifies the global elements exposed to the user, a source schema, which describes the structure of the sources in the system, and a mapping, which specifies the relationship between the sources and the global schema. User queries are posed on the global schema, and the system provides the answers to such queries by exploiting the information supplied by the mapping and accessing the sources that contain relevant data. Thus, from the syntactic viewpoint, the specification of an integration system depends on the following parameters:

- The form of the global schema, i.e., the formalism used for expressing global elements and relationships between global elements, e.g., integrity constraints expressed over a database schema. Several settings have been considered in the literature, where, for instance, the global schema can be relational [13], object-oriented [4], semi-structured [22], based on Description Logics [16, 11], etc..
- The form of the source schema, i.e., the formalism used for expressing data at the sources (as presented by wrappers) and relationships between such data. In principle, the formalisms commonly adopted for the source schema are the same as those mentioned for the global schema.
- The form of the mapping. As already said, two basic approaches have been proposed in the literature, called respectively *global-as-view* (GAV) and *local-as-view* (LAV) [20, 25]. The GAV approach requires that the global schema is defined in terms of the data sources: more precisely, 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. Conversely, in the LAV approach, the meaning of the sources is specified in terms of the elements of the global schema: more precisely, 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.
- The language of the mapping, i.e., the query language used to express views in the mapping.
- The language of the user queries, i.e., the query language adopted by users to issue queries on the global schema.

Let us now turn our attention on the semantics. According to [18], the semantics of a data integration system is given in terms of instances of the global schema (e.g., one set of tuples for each global relation if the global schema is relational, one set of objects for each global class if it is object-oriented, etc.). Such instances have to satisfy *(i)* the knowledge expressed by the global schema, and *(ii)* the mapping specified between the global and the source schema. Roughly speaking, the satisfaction of the mapping depends on the data stored at the sources, and the semantic interpretation of the views in the mapping (see below for more details). Observe that the specification of preferences on source data, if available, contributes to such a notion of satisfaction.

In the following, we give a precise characterization of the concepts informally explained above. In particular, in the line of [18], we provide a logical formal framework which captures all the syntactic and semantic aspects of data integration applications. Here, we consider languages for the specification of the global and the source schema, the mapping and the user queries that rely on First-Order Logic (FOL). Actually, the expressive power of FOL allows us to capture most of the approaches to data integration proposed in the literature. Moreover, in this paper we focus on LAV mappings, which are often considered more appropriate when data sources are autonomous, and may be dynamically changed, added or removed from the data integration system [18].

2.1 Syntax

A data integration system $\mathcal{I}$ is a triple $\langle \mathcal{G}, \mathcal{S}, \mathcal{M} \rangle$, where:

- $\mathcal{G}$ is the *global schema*, which is specified in some subset of FOL on an alphabet formed by a set $\mathcal{A}_\mathcal{G}$ of relation names (or predicates) with associated arity, and a (denumerable) set of constant symbols Γ(we do not consider functions);
- $\mathcal{S}$ is the *source schema*, composed by the schemas of the various sources. We assume that the source schema is simply a set of relation names (with associated arity) of an alphabet $\mathcal{A}_\mathcal{S}$. In other words, we do not allow for the specification of FOL assertions establishing integrity constraints over data sources. This implies that data stored at the sources are always considered locally consistent. This is a common assumption in data integration, because sources are in general autonomous and external to the integration system, which is not in charge to analyze their consistency;
- $\mathcal{M}$ is the *mapping* between $\mathcal{G}$ and $\mathcal{S}$. It is constituted by a set of LAV *assertions* in which, intuitively, views, i.e., queries, expressed over $\mathcal{G}$ are put in correspondence to source relations in $\mathcal{S}$. We assume that queries in the mapping are conjunctive queries, i.e., open formulas of the form

  $$\{x_1, \ldots, x_n \mid \exists y_1, \ldots, y_m . conj(x_1, \ldots, x_n, y_1, \ldots, y_m)\}$$

  where *conj* is a conjunction of atoms, whose predicate symbols are relation names in $\mathcal{A}_\mathcal{G}$, $x_1, \ldots, x_n$ is the sequence of free variables of the query, and n is the *arity* of the query. We omit $\exists y_1, \ldots, y_m$, when clear from the context. More precisely, a mapping assertion assumes the form

  $$s \rightsquigarrow q_g$$

  where s is a relation name of $\mathcal{S}$ and q_g is a conjunctive query over $\mathcal{A}_\mathcal{G}$ of the same arity of s.

Finally, we consider *user queries* posed to a data integration system $\mathcal{I}$, and define their syntax. Each such query q is an open formula that specifies which data to extract from the integration system, i.e., q is intended to extract a set of elements of the domain of interpretation (see below). We assume that user queries are conjunctive queries over the alphabet $\mathcal{A}_\mathcal{G}$ of the global schema $\mathcal{G}$.

EXAMPLE 2.1 Consider a data integration system $\mathcal{I}_0 = \langle \mathcal{G}_0, \mathcal{S}_0, \mathcal{M}_0 \rangle$, where the global schema alphabet $\mathcal{A}_{\mathcal{G}_0}$ comprises the three binary predicates *CourseRoom*, *RoomCapacity* and *SeminarRoom*, which respectively indicate rooms with courses taught there in, how many seats are in the rooms, and rooms in which seminars are scheduled. Assume that the following FOL sentences are specified over the alphabet $\mathcal{A}_{\mathcal{G}_0}$:

$$\forall x, y_1, y_2 . CourseRoom(x, y_1) \wedge CourseRoom(x, y_2) \supset y_1 = y_2$$
$$\forall x_1, y_1, x_2, y_2 . CourseRoom(x_1, y_1) \wedge SeminarRoom(x_2, y_2) \supset y_1 \neq y_2$$
$$\forall x, y . CourseRoom(x, y) \supset \exists z . RoomCapacity(y, z)$$

which state respectively that a course is taught in exactly one room, that seminars and courses are assigned to different rooms, and that the number of seats for each room in which courses are taught is known.

Consider now the source schema $\mathcal{S}_0$, and assume that its alphabet $\mathcal{A}_{\mathcal{S}_0}$ comprises the ternary relation name s_1 and the two binary relation names s_2 and s_3 which provide respectively courses with the room in which they are taught and its capacity, rooms in which courses are taught (but not the capacity of the room), and rooms in which seminars are scheduled.

According to the above description of the sources, we define the mapping $\mathcal{M}_0$ with the following three assertions:

$$s_1 \rightsquigarrow \{x, y, z \mid CourseRoom(x, y) \wedge RoomCapacity(y, z)\}$$
$$s_2 \rightsquigarrow \{x, y \mid CourseRoom(x, y)\}$$
$$s_3 \rightsquigarrow \{x, y \mid SeminarRoom(x, y)\}$$

Finally, we consider the following query issued on the global schema

$$\{x \mid RoomCapacity(x, y)\},$$

that asks for the all rooms for which we know the capacity. ■

2.2 Semantics

We assume that the domain of interpretation is a fixed denumerable set of elements Δ, and that every such element is denoted uniquely by a constant symbol in Γ. In this way, constants in Γ act as *standard names* [19].

Intuitively, to specify the semantics of a data integration system, we have to start with a set of data at the sources, and we have to specify which are the data that satisfy the global schema with respect to such data at the sources. Thus, in order to assign the semantics to a data integration system $\mathcal{I} = \langle \mathcal{G}, \mathcal{S}, \mathcal{M} \rangle$, we start by considering a *source model* for $\mathcal{I}$, i.e., a model $\mathcal{D}$ for the source schema $\mathcal{S}$.

Based on $\mathcal{D}$, we specify the information content of the global schema $\mathcal{G}$. We call any interpretation over Δ of the symbols in $\mathcal{A}_\mathcal{G}$ a *global interpretation* for $\mathcal{I}$.

DEFINITION 2.1. *Let $\mathcal{I} = \langle \mathcal{G}, \mathcal{S}, \mathcal{M} \rangle$ be a data integration system, let $\mathcal{D}$ be a source model for $\mathcal{I}$, a global interpretation $\mathcal{B}$ for $\mathcal{I}$ is a* model for $\mathcal{I}$ w.r.t. $\mathcal{D}$ *if the following conditions hold:*

1. *$\mathcal{B}$ is a model of $\mathcal{G}$, i.e., $\mathcal{B} \models \mathcal{G}$;*
2. *$\mathcal{B}$* satisfies the mapping *$\mathcal{M}$ wrt $\mathcal{D}$.*

Roughly speaking, the notion of satisfying a LAV mapping depends on

(*a*) criteria adopted to deal with inconsistency, and

(*b*) criteria used to interpret source preferences, if available.

We will use the symbol X to specify a certain semantic criterion characterizing both point (*a*) and (*b*) above.

DEFINITION 2.2. *Let $\mathcal{I} = \langle \mathcal{G}, \mathcal{S}, \mathcal{M} \rangle$ be a data integration system, let $\mathcal{D}$ be a source model for $\mathcal{I}$, and let X be a semantic criterion. The set of all models for $\mathcal{I}$ w.r.t. $\mathcal{D}$, denoted by $sem_X(\mathcal{I}, \mathcal{D})$, is called the* semantics of $\mathcal{I}$ w.r.t. $\mathcal{D}$ under X.

In the following, in place of X we will use a different subscript for each criterion which we consider in this paper.

Let us now turn our attention to queries. In order to define the semantics of a query q over a data integration system $\mathcal{I}$, we have to take into account all the interpretations of $\mathcal{G}$ in the semantics of $\mathcal{I}$ with respect to $\mathcal{D}$.

DEFINITION 2.3. *Let $\mathcal{I} = \langle \mathcal{G}, \mathcal{S}, \mathcal{M} \rangle$ be a data integration system, let $\mathcal{D}$ be a source model for $\mathcal{I}$, let X be a semantic criterion, and let q be a user query of arity n over $\mathcal{I}$. The set of* certain answers *of q with respect to $\mathcal{I}$ and $\mathcal{D}$ under X, denoted by $ans_X(q, \mathcal{I}, \mathcal{D})$, is defined as follows:*

$$ans_X(q, \mathcal{I}, \mathcal{D}) = \{\langle c_1, \ldots, c_n \rangle \mid \text{for each } \mathcal{B} \in sem_X(\mathcal{I}, \mathcal{D}), \langle c_1, \ldots, c_n \rangle \in q^{\mathcal{B}} \}$$

where $q^{\mathcal{B}}$ denotes the result of evaluating q in the interpretation $\mathcal{B}$, i.e., the set of n-tuples of elements of Δ associated to the free variables of q (recall that the interpretation of constants, i.e., standard names, is the same in every interpretation).

Such a notion of answers, corresponding to skeptical entailment, is the most used in data integration [15, 18]; however the notion of *possible answers*, corresponding to credulous entailment, can also be defined.

3. FOL SEMANTICS

In this section, we consider the case in which no preferences are specified over the sources, and adopt a classical FOL interpretation of the mapping. According to such an interpretation, the mapping can be exploited in order to infer extensions of the global schema starting from the data stored at the sources. In other words, data at the sources form a partial specification of the intended models of the theory provided by the integration system. In this situation, computing the certain answers to a user query is an inference process similar to query answering in the presence of incomplete information.

DEFINITION 3.1. *Let $\mathcal{I} = \langle \mathcal{G}, \mathcal{S}, \mathcal{M} \rangle$ be an integration system, let $\mathcal{D}$ be a source model for $\mathcal{I}$, and let $\mathcal{B}$ be a global interpretation for $\mathcal{I}$. Then, $\mathcal{B}$* satisfies (the FOL interpretation of) *$\mathcal{M}$ with respect to $\mathcal{D}$ if, for each assertion in $\mathcal{M}$ of the form $s \rightsquigarrow q_g$, it holds*

$$s^{\mathcal{D}} \subseteq q_g^{\mathcal{B}},$$

where $s^{\mathcal{D}}$ denotes the evaluation of s in $\mathcal{D}$, i.e., the set of tuples of elements of Δ (i.e., standard names) assigned to s by $\mathcal{D}$, and $q_g^{\mathcal{B}}$ denotes the evaluation of q_g over $\mathcal{B}$. In other words, an assertion of the form $s \rightsquigarrow q_g$ is satisfied if each tuple in $s^{\mathcal{D}}$ is also a tuple of $q_g^{\mathcal{B}}$.

The above interpretation of the mapping is also called *sound* interpretation in the literature [18], and is commonly adopted in data integration, for being it able to capture the incomplete nature of data sources w.r.t. the intended extension of the global schema, which is a common setting in data integration.

With this notion in place we can provide the *FOL semantics* of data integration system $\mathcal{I} = \langle \mathcal{G}, \mathcal{S}, \mathcal{M} \rangle$ w.r.t. a source model $\mathcal{D}$ for $\mathcal{I}$, denoted $sem_{FOL}(\mathcal{I}, \mathcal{D})$, which, according to Definition 2.2, is the set of global interpretations $\mathcal{B}$ for $\mathcal{I}$ such that

1. $\mathcal{B} \models \mathcal{G}$;
2. *$\mathcal{B}$ satisfies the FOL interpretation of the mapping $\mathcal{M}$* wrt *$\mathcal{D}$.*

Analogously, by applying definition 2.3 to the FOL semantics, we characterize the certain answers to a user query q under sem_{FOL}, which we denote by $ans_{FOL}(q, \mathcal{I}, \mathcal{D})$.

EXAMPLE 2.1 (CONTD.) Assume now that the domain Δ contains, among others, the elements Analysis, Geometry, A1, A2, A3, Data Quality, 215, and let $\mathcal{D}_0$ be a source model for $\mathcal{I}_0$ such that the set of facts that hold in $\mathcal{D}_0$ is as follows:

$\{s_1(\text{Analysis}, \text{A1}, 215), s_2(\text{Geometry}, \text{A2}), s_3(\text{Data Quality}, \text{A3})\}$.

Consider the following set $\mathcal{B}_0$ of facts

$\{CourseRoom(\text{Analysis}, \text{A1}), RoomCapacity(\text{A1}, 215), CourseRoom(\text{Geometry}, \text{A2}), SeminarRoom(\text{Data Quality}, \text{A1}), RoomCapacity(\text{A2}, \alpha)\}$.

where α is a constant of the domain Δ.

It is easy to see that $\mathcal{B}_0 \models \mathcal{G}_0$, and that, modulo the choice of α from Δ, the above set of facts holds for all FOL models of $\mathcal{I}_0$ with respect to $\mathcal{D}_0$, i.e., $sem_{FOL} = \{\mathcal{B} \mid \mathcal{B} \models \mathcal{G} \text{ and } \mathcal{B} \supseteq \mathcal{B}_0\}$. Hence, for the query $q = \{x \mid RoomCapacity(x, y)\}$ we have that $ans_{FOL}(q, \mathcal{I}_0, \mathcal{D}_0) = \{\text{A1}, \text{A2}\}$. ∎

4. MAXIMALLY-SOUND SEMANTICS

According to the semantics $sem_{FOL}(\mathcal{I}, \mathcal{D})$, it may be the case that the data retrieved from the sources cannot be reconciled in the global schema in such a way that both the knowledge in the global schema and the mapping are satisfied [17]. This is in general caused by mutual inconsistencies in the data coming from different sources. In such cases, $sem_{FOL}(\mathcal{I}, \mathcal{D}) = \emptyset$, therefore, by Definition 2.3, every n-tuple is in the answer set of every query of arity n. This is not an acceptable way of handling inconsistency: as motivated by the studies in consistent query answering in inconsistent databases [8, 2, 14], it could be possible to derive significant answers to queries even in the presence of inconsistency.

EXAMPLE 2.1 (CONTD.) Suppose now to have a different source model $\mathcal{D}_1$ such that the following set of facts hold:

$\{s_1(\text{Analysis}, \text{A1}, 215), s_2(\text{Analysis}, \text{A2}), s_3(\text{Data Quality}, \text{A2})\}$.

It should be easy to see that in this case $sem(\mathcal{I}_0, \mathcal{D}_1) = \emptyset$. Indeed, according to the mapping assertions, for each global interpretation that satisfies $\mathcal{M}_0$, both the facts $CourseRoom(\text{Analysis}, \text{A1})$ and $CourseRoom(\text{Analysis}, \text{A2})$ hold. On the other hand, such facts together violate the assertion of $\mathcal{G}_0$ stating that a course is taught only in a room. Furthermore, from data stored in s_3 we also infer on the global schema the fact $SeminarRoom(\text{Data Quality}, \text{A2})$, which together with the fact $CourseRoom(\text{Analysis}, \text{A2})$ violates the assertion stating that rooms assigned to courses and seminars have to be different.

Hence the system $\mathcal{I}_0$ is inconsistent with respect to the source model $\mathcal{D}_1$, and the certain answers to each query of arity n are all the n-tuples of elements of Δ. Nonetheless, in each interpretation that satisfies $\mathcal{M}_0$ w.r.t. $\mathcal{D}_1$ we have that Analysis is a course. Intuitively, we would preserve this

knowledge even in the presence of inconsistencies that do not directly contradict the fact that Analysis is a course. Consider for example the user query $\{x \mid CourseRoom(x, y)\}$: is reasonable to assume that the set of "significant" certain answers to this query is the set {Analysis}, rather than the entire domain Δ. ∎

To the aim of overcoming the problems illustrated above, we introduce a different notion of mapping satisfaction, which can be intuitively seen as a relaxation of the FOL interpretation discussed in Section 3. In other words, we adopt a different criterion to deal with inconsistent data[1]. More precisely, global interpretations of a data integration system $\mathcal{I} = \langle \mathcal{G}, \mathcal{S}, \mathcal{M} \rangle$ that now we are looking for, are those interpretations that satisfy $\mathcal{G}$ and that satisfy *as much as possible* the (FOL interpretation of the) mapping assertions in $\mathcal{M}$ w.r.t. a source model $\mathcal{D}$ for $\mathcal{I}$. In other words, in our approach, the knowledge expressed by $\mathcal{G}$ is considered more reliable than the knowledge represented by the information retrieved at the data sources through the mapping assertions.

In order to determine the precise meaning of "satisfying as much as possible" the mapping with respect to a source model $\mathcal{D}$, we define preference orders over the models of $\mathcal{G}$. Informally, we consider as intended models of the integration system those models of $\mathcal{G}$ that satisfy as much as possible a set of first-order sentences that constitutes the "image of the mapping assertions" with respect to $\mathcal{D}$.

To formalize the above ideas, we first define the notions of "image" of the mapping $\mathcal{M}$ with respect to a model $\mathcal{D}$ of the sources as a set of first-order sentences. In the following definition, $q(t)$ indicates the FOL sentence obtained from the open formula q by replacing its free variables with the constants in t, i.e., if $t = \{t_1, \ldots, t_n\}$ and $\{x_1, \ldots, x_n\}$ are the free variables of q, $x_i = t_i$ for each $1 \leq i \leq n$.

DEFINITION 4.1. *Let $\mathcal{I} = \langle \mathcal{G}, \mathcal{S}, \mathcal{M} \rangle$ be a data integration system, and $\mathcal{D}$ a source model for $\mathcal{I}$, we define the following set of FOL sentences $Image(\mathcal{M}, \mathcal{D})$:*

$$Image(\mathcal{M}, \mathcal{D}) \quad = \quad \{q_g(t) \mid s \rightsquigarrow q_g \in \mathcal{M} \text{ and } t \in s^{\mathcal{D}}\}.$$

Roughly speaking, $Image(\mathcal{M}, \mathcal{D})$ contains all the FOL sentences that are implied by the source model $\mathcal{D}$ and the mapping assertions in $\mathcal{M}$. In other words, $Image(\mathcal{M}, \mathcal{D})$ is the minimal set of FOL sentences that hold at the global level for a source model $\mathcal{D}$.

EXAMPLE 2.1 (CONTD.) In our ongoing example, we have

$$\begin{aligned} Image(\mathcal{M}_0, \mathcal{D}_1) = \{ & \\ & CourseRoom(\mathsf{Analysis}, \mathsf{A1}) \wedge RoomCapacity(\mathsf{A1}, 215), \\ & CourseRoom(\mathsf{Analysis}, \mathsf{A2}), \\ & SeminarRoom(\mathsf{Data\ Quality}, \mathsf{A2})\}. \end{aligned}$$

∎

Then, given an interpretation $\mathcal{W}$ of the global schema $\mathcal{G}$, we define $SatImage(\mathcal{W}, \mathcal{M}, \mathcal{D})$ as the portion of the image of $\mathcal{M}$ with respect to $\mathcal{D}$ satisfied by $\mathcal{W}$. More precisely:

[1]Note that also in this section we do not consider preferences specified on data sources.

DEFINITION 4.2. *let $\mathcal{I} = \langle \mathcal{G}, \mathcal{S}, \mathcal{M} \rangle$ be a data integration system, let $\mathcal{D}$ be a source model for $\mathcal{I}$, and let $\mathcal{W}$ be a global interpretation of $\mathcal{I}$. We define:*

$$SatImage(\mathcal{W}, \mathcal{M}, \mathcal{D}) = \{\varphi \mid \varphi \in Image(\mathcal{M}, \mathcal{D}) \text{ and } \mathcal{W} \models \varphi\}.$$

Based on the above notions of image of the mapping with respect to a source model, we now define a partial order (based on set containment) over the interpretations of the global schema.

DEFINITION 4.3. *Let $\mathcal{I} = \langle \mathcal{G}, \mathcal{S}, \mathcal{M} \rangle$ be a data integration system, let $\mathcal{D}$ be a source model for $\mathcal{I}$, and let $\mathcal{W}, \mathcal{W}'$ be two global interpretations of $\mathcal{I}$. We say that $\mathcal{W}'$ is $(\mathcal{M}, \mathcal{D})$*-preferred to $\mathcal{W}$ *if $SatImage(\mathcal{W}', \mathcal{M}, \mathcal{D}) \supset SatImage(\mathcal{W}, \mathcal{M}, \mathcal{D})$.*

Then, we are ready to generalize Definition 3.1 and give a new notion of global models that satisfies the mapping, which corresponds to the notion of maximal element in the partial order defined above.

DEFINITION 4.4. *Let $\mathcal{I} = \langle \mathcal{G}, \mathcal{S}, \mathcal{M} \rangle$ be a data integration system, let $\mathcal{D}$ be a source model for $\mathcal{I}$, and let $\mathcal{W}$ be a model of $\mathcal{G}$. We say that $\mathcal{W}$* maximally satisfies *$\mathcal{M}$ if for each model $\mathcal{W}'$ of $\mathcal{G}$, $\mathcal{W}'$ is not $(\mathcal{M}, \mathcal{D})$-preferred to $\mathcal{W}$.*

It is easy now to define the *maximally-sound semantics* of a data integration system $\mathcal{I} = \langle \mathcal{G}, \mathcal{S}, \mathcal{M} \rangle$ w.r.t. a source model $\mathcal{D}$ for $\mathcal{I}$, denoted $sem_{MS}(\mathcal{I}, \mathcal{D})$, which is actually the set of global interpretations $\mathcal{W}$ for $\mathcal{I}$ such that

1. $\mathcal{W} \models \mathcal{G}$;
2. $\mathcal{W}$ *maximally satisfies* $\mathcal{M}$ wrt $\mathcal{D}$.

The certain answers to a user query q under the maximally-sound semantics are denoted by $ans_{MS}(q, \mathcal{I}, \mathcal{D})$.

EXAMPLE 2.1 (CONTD.) Let us first enumerate the sentences of $Image(\mathcal{M}_0, \mathcal{D}_1)$ as follows:

1. $CourseRoom(\mathsf{Analysis}, \mathsf{A1}) \wedge RoomCapacity(\mathsf{A1}, 215)$;
2. $CourseRoom(\mathsf{Analysis}, \mathsf{A2})$;
3. $SeminarRoom(\mathsf{Data\ Quality}, \mathsf{A2})$.

Then, according to the above definitions, we have that $sem_{MS}(\mathcal{I}_0, \mathcal{D}_1)$ contains all models $\mathcal{W}$ for $\mathcal{G}$ such that they satisfy either sentences 1 and 3 or sentence 2. Indeed, if $\mathcal{W}$ satisfied all sentences in $Image(\mathcal{M}_0, \mathcal{D}_1)$, as already noticed, the facts $CourseRoom(\mathsf{Analysis}, \mathsf{A1})$ and $CourseRoom(\mathsf{Analysis}, \mathsf{A2})$ would hold in $\mathcal{W}$, thus violating the assertion in $\mathcal{G}_0$ stating that each course is taught exactly in one room. Analogously, the facts $CourseRoom(\mathsf{Analysis}, \mathsf{A2})$ and $SeminarRoom(\mathsf{Data\ Quality}, \mathsf{A2})$ would violated the asserted separation between seminar and course rooms. On the other hand, $\mathcal{W}$ cannot satisfy any sentence in $Image(\mathcal{M}_0, \mathcal{D}_1)$, since in such a way it would not be maximal w.r.t. the $(\mathcal{M}_0, \mathcal{D}_1)$-preference ordering. ∎

We point out that the semantics sem_{MS} defined above has an important property: for each integration system $\mathcal{I}$ and source model $\mathcal{D}$, if $sem_{FOL}(\mathcal{I},\mathcal{D}) \neq \emptyset$ then $sem_{MS}(\mathcal{I},\mathcal{D}) = sem_{FOL}(\mathcal{I},\mathcal{D})$. In this sense, such semantics can be considered as a "conservative extension" of the classical semantics sem_{FOL}, since it provides a different meaning to a data integration system only in the presence of inconsistency (i.e., only when $sem_{FOL}(\mathcal{I},\mathcal{D}) = \emptyset$).

5. ADDING SOURCE PREFERENCES

In this section we consider preferences defined over the sources, and their impact on the notion of mapping satisfaction. Our aim is to provide a generalization of the maximally-sound semantics that allows us to select, among the maximally-sound models, only those that "trust" the best sources. Let us first formally define the notion of preference between two source relations.

DEFINITION 5.1. *Let $\mathcal{I} = \langle \mathcal{G}, \mathcal{S}, \mathcal{M} \rangle$ be a data integration system, a* preference assertion over $\mathcal{S}$ *is an expression of the form*

$$s_i > s_j$$

where s_i, s_j are relation names in $\mathcal{A}_\mathcal{S}$.

The intuitive meaning of preference assertions is that the quality degree of the source relation s_i is higher than that of the source s_j. Notice that, since each source is associated with a mapping assertion, preferences between sources correspond to preferences between mapping assertions, i.e., if $m_i = s_i \rightsquigarrow q_{g_i}$ and $m_j = s_i \rightsquigarrow q_{g_j}$ belong to $\mathcal{M}$, the assertion $s_i > s_j$ implies that m_i is preferred to m_j.

Then, a data integration system with preference assertions is a four-tuple of the form $\langle \mathcal{G}, \mathcal{S}, \mathcal{M}, \mathcal{P} \rangle$, where $\mathcal{G}$, $\mathcal{S}$, $\mathcal{M}$ are as before and $\mathcal{P}$ is a set of preference assertions over $\mathcal{S}$.

In the following, we denote by $\mathcal{P}^*$ the set of assertions representing the transitive closure of the binary relation $>$ in $\mathcal{P}$, i.e., $\mathcal{P}^*$ is the least set of assertions such that: (i) $\mathcal{P}^* \supseteq \mathcal{P}$; (ii) if $s_1 > s_2 \in \mathcal{P}^*$ and $s_2 > s_3 \in \mathcal{P}^*$, then $s_1 > s_3 \in \mathcal{P}^*$.

DEFINITION 5.2. *Let $\mathcal{I} = \langle \mathcal{G}, \mathcal{S}, \mathcal{M}, \mathcal{P} \rangle$ be a data integration system with preference assertions, let $\mathcal{D}$ be a source model for $\mathcal{I}$, and let $\mathcal{W}$,$\mathcal{W}'$ be two interpretations of $\mathcal{G}$. Then, we write $SatImage(\mathcal{W}', \mathcal{M}, \mathcal{D}) >_\mathcal{P} SatImage(\mathcal{W}, \mathcal{M}, \mathcal{D})$ if*

(a) there exists a preference assertion $s_1 > s_2 \in \mathcal{P}^*$, with $s_1 \rightsquigarrow q_{g_1} \in \mathcal{M}$, $s_2 \rightsquigarrow q_{g_2} \in \mathcal{M}$, such that there exist $q_{g_1}(t_1) \in SatImage(\mathcal{W}', \mathcal{M}, \mathcal{D}) - SatImage(\mathcal{W}, \mathcal{M}, \mathcal{D})$ and $q_{g_2}(t_2) \in SatImage(\mathcal{W}, \mathcal{M}, \mathcal{D}) - SatImage(\mathcal{W}', \mathcal{M}, \mathcal{D})$, and

(a) there exists no preference assertion $s_0 > s_1 \in \mathcal{P}^*$ with $s_0 \rightsquigarrow q_{g_0} \in \mathcal{M}$, such that there exists $q_{g_0}(t_0) \in SatImage(\mathcal{W}, \mathcal{M}, \mathcal{D}) - SatImage(\mathcal{W}', \mathcal{M}, \mathcal{D})$.

If $SatImage(\mathcal{W}', \mathcal{M}, \mathcal{D}) >_\mathcal{P} SatImage(\mathcal{W}, \mathcal{M}, \mathcal{D})$ does not hold, then we write $SatImage(\mathcal{W}', \mathcal{M}, \mathcal{D}) \not>_\mathcal{P} SatImage(\mathcal{W}, \mathcal{M}, \mathcal{D})$.

Roughly speaking, condition (a) of the above definition imposes that $SatImage(\mathcal{W}', \mathcal{M}, \mathcal{D})$ contains a sentence $(q_{g_1}(t_1))$ that is not contained in $SatImage(\mathcal{W}, \mathcal{M}, \mathcal{D})$ and that is preferable to a sentence $(q_{g_2}(t_2))$ in $SatImage(\mathcal{W}, \mathcal{M}, \mathcal{D})$. Furthermore, condition (b) imposes that there does not exist a sentence $(q_{g_0}(t_0))$ in $SatImage(\mathcal{W}, \mathcal{M}, \mathcal{D})$, and not contained in $SatImage(\mathcal{W}', \mathcal{M}, \mathcal{D})$, that is in turn preferable to $q_{g_1}(t_1)$. Intuitively, condition (b) guarantees that $SatImage(\mathcal{W}', \mathcal{M}, \mathcal{D}) >_\mathcal{P} SatImage(\mathcal{W}, \mathcal{M}, \mathcal{D})$ only if there exists a sentence in $SatImage(\mathcal{W}', \mathcal{M}, \mathcal{D})$ that is never "worse" than a sentence of $SatImage(\mathcal{W}, \mathcal{M}, \mathcal{D})$.

With this notion in place, we can define a preference ordering between the models of a data integration system.

DEFINITION 5.3. *Let $\mathcal{I} = \langle \mathcal{G}, \mathcal{S}, \mathcal{M}, \mathcal{P} \rangle$ be a data integration system with source preferences, let $\mathcal{D}$ be a source model of $\mathcal{I}$, and let $\mathcal{W}, \mathcal{W}'$ be two interpretations of $\mathcal{G}$ such that $\mathcal{W} \models \mathcal{G}$, $\mathcal{W}' \models \mathcal{G}$. We say that $\mathcal{W}'$ is* $(\mathcal{M}, \mathcal{D}, \mathcal{P})$-preferred *to $\mathcal{W}$ if at least one of the following conditions holds:*

1. *$SatImage(\mathcal{W}', \mathcal{M}, \mathcal{D}) \supset SatImage(\mathcal{W}, \mathcal{M}, \mathcal{D})$;*
2. *$SatImage(\mathcal{W}', \mathcal{M}, \mathcal{D}) >_\mathcal{P} SatImage(\mathcal{W}, \mathcal{M}, \mathcal{D})$ and $SatImage(\mathcal{W}, \mathcal{M}, \mathcal{D}) \not>_\mathcal{P} SatImage(\mathcal{W}', \mathcal{M}, \mathcal{D})$.*

Then, we say that $\mathcal{W}$ maximally satisfies *$(\mathcal{M}, \mathcal{P})$ w.r.t. $\mathcal{D}$ if $\mathcal{W}$ is a model for $\mathcal{G}$, and for each model $\mathcal{W}'$ for $\mathcal{G}$, $\mathcal{W}'$ is not $(\mathcal{M}, \mathcal{D}, \mathcal{P})$-preferred to $\mathcal{W}$.*

According to the above definition, the *preference-based semantics* of a data integration system $\mathcal{I} = \langle \mathcal{G}, \mathcal{S}, \mathcal{M}, \mathcal{P} \rangle$ w.r.t. a source model $\mathcal{D}$ for $\mathcal{I}$, denoted $sem_{MSP}(\mathcal{I},\mathcal{D})$, is the set of global interpretations $\mathcal{W}$ for $\mathcal{I}$ such that

1. $\mathcal{W} \models \mathcal{G}$;
2. $\mathcal{W}$ maximally satisfies $(\mathcal{M}, \mathcal{P})$ w.r.t. $\mathcal{D}$.

The certain answers to a user query q under the preferred semantics are denoted by $ans_{MSP}(\mathcal{I},\mathcal{D})$.

It is easy to show that, given an integration system with source preferences $\mathcal{I} = \langle \mathcal{G}, \mathcal{S}, \mathcal{M}, \mathcal{P} \rangle$, if we call $\mathcal{I}' = \langle \mathcal{G}, \mathcal{S}, \mathcal{M} \rangle$ the integration system without source preferences obtained from $\mathcal{I}$, then, for each source model $\mathcal{D}$ for $\mathcal{S}$, $sem_{MSP}(\mathcal{I},\mathcal{D}) \subseteq sem_{MS}(\mathcal{I}',\mathcal{D})$.

EXAMPLE 2.1 (CONTD.) Let us recall the mapping $\mathcal{M}_0$ of the integration system $\mathcal{I}_0$:

$s_1 \rightsquigarrow \{x, y, z \mid CourseRoom(x, y) \wedge RoomCapacity(y, z)\}$
$s_2 \rightsquigarrow \{x, y \mid CourseRoom(x, y)\}$
$s_3 \rightsquigarrow \{x, y \mid SeminarRoom(x, y)\}$

Then, let $\mathcal{P}_0$ be the following set of preferences:

$$\mathcal{P}_0 = \{s_1 > s_2,\ s_2 > s_3\}$$

which express that the quality of the source relation s_1 is higher than the quality of the source relation s_2, which in turn is preferred to source relation s_3. Hence, we easily obtain $\mathcal{P}_0^* = \mathcal{P}_0 \cup \{s_1 > s_3\}$.

Let us also recall enumeration of the sentences of $Image(\mathcal{M}_0, \mathcal{D}_1)$ previously introduced:

1. $CourseRoom(\mathsf{Analysis}, \mathsf{A1}) \wedge RoomCapacity(\mathsf{A1}, 215)$;
2. $CourseRoom(\mathsf{Analysis}, \mathsf{A2})$;
3. $SeminarRoom(\mathsf{Data\ Quality}, \mathsf{A2})$.

Finally, consider $\mathcal{I}_1 = \langle \mathcal{G}_0, \mathcal{S}_0, \mathcal{M}_0, \mathcal{P}_0 \rangle$. By applying the above Definition 5.3, we have that $sem_{MSP}(\mathcal{I}_1, \mathcal{D}_1)$ is constituted by the models of $sem_{MS}(\mathcal{I}_0, \mathcal{D}_1)$ that satisfy sentences 1 and 3. Such interpretations are the ones that reflect the preference ordering over the sources given by the assertions in $\mathcal{P}_0^*$. Hence, for the query $q = \{x, y \mid CourseRoom(x, y)\}$ we have that $ans_{MSP}(q, \mathcal{I}_1, \mathcal{D}_1) = \{\langle \mathsf{Analysis}, \mathsf{A1} \rangle\}$, while $ans_{MS}(q, \mathcal{I}_0, \mathcal{D}_1) = \emptyset$,. ∎

6. CONCLUSIONS

In this paper, we have laid the semantic foundations for a LAV data integration system whose declarative specification includes information about source preferences. The semantics that we propose is the first one to deal both with information on source reliability and with inconsistent data in LAV data integration systems.

We believe that the approach presented in this paper may be extended in several ways. First, we can easily extend the approach to more expressive forms of (sound) mapping, e.g., GLAV mapping, a generalization of GAV and LAV [18]. Also, it should be noted that the treatment of preferences proposed here can be refined to give preferences at the tuple-level, instead of at the whole source level. This extension is particularly interesting since in principle it allows to assign context-dependent preferences, where each context is represented by an appropriate query at the sources.

Finally we mention that the computational aspects of the framework need to be investigated: sound, complete and terminating techniques for query answering in data integration systems with source preferences have to be defined. In doing so, particular attention must be posed to single out those cases that allow for efficient (i.e., polynomially bounded) computation on the data.

7. ACKNOWLEDGMENTS

This research has been partially supported by the projects INFOMIX (IST-2001-33570), SEWASIE (IST-2001-34825) and INTEROP Network of Excellence (IST-508011) funded by the EU, by the project "Società dell'Informazione" subproject SP1 "Reti Internet: Efficienza, Integrazione e Sicurezza" funded by MIUR – Fondo Speciale per lo Sviluppo della Ricerca di Interesse Strategico, and by project HYPER, funded by IBM through a Shared University Research (SUR) Award grant.

8. REFERENCES

[1] Philipp Anokhin and Amihai Motro. Fusionplex: Resolution of data inconsistencies in the integration of heterogeneous inofmration sources. Technical Report ISE-TR-03-06, Department of Information and Software Engineering, George Mason University, 2003.

[2] Marcelo Arenas, Leopoldo E. Bertossi, and Jan Chomicki. Consistent query answers in inconsistent databases. In *Proc. of the 18th ACM SIGACT SIGMOD SIGART Symp. on Principles of Database Systems (PODS'99)*, pages 68–79, 1999.

[3] Marcelo Arenas, Leopoldo E. Bertossi, and Jan Chomicki. Specifying and querying database repairs using logic programs with exceptions. In *Proc. of the 4th Int. Conf. on Flexible Query Answering Systems (FQAS 2000)*, pages 27–41. Springer, 2000.

[4] 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.

[5] Leopoldo Bertossi, Jan Chomicki, Alvaro Cortes, and C. Gutierrez. Consistent answers from integrated data sources. In *Proc. of the 6th Int. Conf. on Flexible Query Answering Systems (FQAS 2002)*, pages 71–85, 2002.

[6] Mokrane Bouzeghoub and Maurizio Lenzerini. Introduction to the special issue on data extraction, cleaning, and reconciliation. *Information Systems*, 26(8):535–536, 2001.

[7] 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.

[8] François Bry. Query answering in information systems with integrity constraints. In *IFIP WG 11.5 Working Conf. on Integrity and Control in Information System.* Chapman & Hall, 1997.

[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] Diego Calvanese, Giuseppe De Giacomo, and Maurizio Lenzerini. Answering queries using views over description logics knowledge bases. In *Proc. of the 17th Nat. Conf. on Artificial Intelligence (AAAI 2000)*, pages 386–391, 2000.

[12] Thomas Eiter, Michael Fink, Gianluigi Greco, and Domenico Lembo. Efficient evaluation of logic programs for querying data integration systems. In *Proc. of the 19th Int. Conf. on Logic Programming (ICLP 2003)*, pages 163–177, 2003.

[13] Michael R. Genereseth, Arthur M. Keller, and Oliver M. Duschka. Infomaster: An information integration system. In *ACM SIGMOD International Conference on Management of Data*, 1997.

[14] 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.

[15] Alon Y. Halevy. Answering queries using views: A survey. *Very Large Database J.*, 10(4):270–294, 2001.

[16] Thomas Kirk, Alon Y. Levy, Yehoshua Sagiv, and Divesh Srivastava. The Information Manifold. In *Proceedings of the AAAI 1995 Spring Symp. on Information Gathering from Heterogeneous, Distributed Enviroments*, pages 85–91, 1995.

[17] Domenico Lembo, Maurizio Lenzerini, and Riccardo Rosati. Source inconsistency and incompleteness in data integration. In *Proc. of the 9th Int. Workshop on*

Knowledge Representation meets Databases (KRDB 2002). CEUR Electronic Workshop Proceedings, `http://ceur-ws.org/Vol-54/`, 2002.

[18] 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.

[19] Hector J. Levesque and Gerhard Lakemeyer. *The Logic of Knowledge Bases*. The MIT Press, 2001.

[20] Alon Y. Levy. Logic-based techniques in data integration. In Jack Minker, editor, *Logic Based Artificial Intelligence*. Kluwer Academic Publisher, 2000.

[21] Jinxin Lin and Alberto O. Mendelzon. Merging databases under constraints. *Int. J. of Cooperative Information Systems*, 7(1):55–76, 1998.

[22] Ioana Manolescu, Daniela Florescu, and Donald Kossmann. Answering XML queries on heterogeneous data sources. In *Proc. of the 27th Int. Conf. on Very Large Data Bases (VLDB 2001)*, pages 241–250, 2001.

[23] Felix Naumann, Ulf Leser, and Johann Christoph Freytag. Quality-driven integration of heterogenous information systems. In *Proc. of the 25th Int. Conf. on Very Large Data Bases (VLDB'99)*, pages 447–458, 1999.

[24] Yannis Papakonstantinou, Serge Abiteboul, and Hector Garcia-Molina. Object fusion in mediator systems. In T. M. Vijayaraman, Alejandro P. Buchmann, C. Mohan, and Nandlal L. Sarda, editors, *Proc. of the 22nd Int. Conf. on Very Large Data Bases (VLDB'96)*, pages 413–424, 1996.

[25] Jeffrey D. Ullman. Information integration using logical views. *Theoretical Computer Science*, 239(2):189–210, 2000.

[26] Ling-Ling Yan and M. Tamer Özsu. Conflict tolerant queries in AURORA. In *Proc. of the 7th Int. Conf. on Cooperative Information Systems (CoopIS'99)*, pages 279–290, 1999.