The MOMIS-STASIS approach for Ontology-Based Data Integration

Domenico Beneventano    Mirko Orsini    Laura Po    Serena Sorrentino

Dipartimento di Ingegneria dell’Informazione, Università di Modena e Reggio Emilia, Italy
{firstname.lastname}@unimore.it

Abstract. Ontology based Data Integration involves the use of ontology(s) to effectively combine data and information from multiple heterogeneous sources [18]. Ontologies can be used in an integration task to describe the semantics of the information sources and to make the contents explicit. With respect to the integration of data sources, they can be used for the identification and association of semantically corresponding information concepts, i.e. for the definition of semantic mapping among concepts of the information sources. MOMIS is a Data Integration System which performs information extraction and integration from both structured and semi-structured data sources [7]. The goal of the STASIS project is to create a comprehensive application suite which allows enterprises to simplify the mapping process between data schemas based on semantics [1]. Moreover, in STASIS, a general framework to perform Ontology-driven Semantic Mapping has been proposed [4]. This paper describes the early effort to combine the MOMIS and the STASIS frameworks in order to obtain an effective approach for Ontology-Based Data Integration.

1 Introduction

Data integration is the problem of combining data residing at different autonomous sources, and providing the user with a unified view of these data. The problem of designing Data Integration Systems is important in current real world applications, and is characterized by a number of issues that are interesting from a theoretical point of view [12]. Integration System are usually characterized by a classical wrapper/mediator architecture [19] based on a Global Virtual Schema (Global Virtual View - GVV) and a set of data sources. The data sources contain the real data, while the GVV provides a reconciled, integrated, and virtual view of the underlying sources. Modeling the mappings among sources and the GVV is a crucial aspect. Two basic approaches for specifying the mapping in a Data Integration System have been proposed in the literature: Local-As-View (LAV), and Global-As-View (GAV), respectively [11, 17].

MOMIS (Mediator EnviRonment for Multiple Information Sources) is a Data Integration System which performs information extraction and integration from both structured and semi-structured data sources by following the GAV approach [7, 6]. Information integration is performed in a semi-automatic way, by exploiting the knowledge in a Common Thesaurus (defined by the framework) and descriptions of source schemas with a combination of clustering techniques and Description Logics. This integration
process gives rise to a virtual integrated view of the underlying sources for which mapping rules and integrity constraints are specified to handle heterogeneity.

Ontologies can be used in an integration task to describe the semantics of the information sources and to make the contents explicit [18]. With respect to the integration of data sources, they can be used for the identification and association of semantically corresponding information concepts. In [18], three different approaches of how to employ the ontologies for the explicit description of the information source semantics are identified: single ontology approaches, multiple ontologies approaches and hybrid approaches. Single ontology approaches use one global ontology providing a shared vocabulary for the specification of the semantics: all data sources are related to one global ontology. In multiple ontology approaches, each information source is described by its own ontology and mappings between the ontologies are defined: these inter-ontology mappings identify semantically corresponding terms of different source ontologies, e.g. which terms are semantically equal or similar. In hybrid approaches similar to multiple ontology approaches the semantics of each source is described by its own ontology, but in order to make the source ontologies comparable to each other they are built upon one global shared vocabulary which contains basic terms of a domain [18].

With respect to the above classification, the MOMIS Data Integration System uses a single ontology approach, where the lexical ontology WordNet [14] is used as a shared vocabulary for the specification of the semantics of data sources and for the identification and association of semantically corresponding information concepts. The main reason of this choice is that, by using a lexical ontology as WordNet (which it is characterized by a wide network of semantic relationships between concepts), the annotation of data sources elements can be performed in a semi-automatic way by using Word Sense Disambiguation techniques.

The STASIS IST project (www.stasis-project.net) is a Research and Development project sponsored under the EC 6th Framework programme. It aims to enable SMEs and enterprises to fully participate in the Economy, by offering semantic services and applications based on the open SEEM registry and repository network. The goal of the STASIS project is to create a comprehensive application suite which allows enterprises to simplify the mapping process between data schemas, by providing an easy to use GUI, allowing users to identify semantic elements in an easy way [2, 1].

Moreover, in the STASIS project, a general framework to perform Ontology-driven Semantic Mapping has been proposed, where the identification of mappings between concepts of different schemas is based on the schemas annotation with respect to ontologies [4].

In [5] this framework has been further elaborated and it has been applied to the context of products and services catalogues. In the STASIS project OWL is used as language to include in the framework generic external ontologies.

This paper describes an approach to combine the MOMIS and STASIS frameworks in order to obtain an effective approach for Ontology-Based Data Integration. The proposal is based on the extension of the MOMIS system by using the Ontology-driven Semantic Mapping framework developed in STASIS in order to address the following points:
1. enabling the MOMIS system to employ generic OWL ontologies, with respect to the limitation of using only the WordNet lexical ontology;
2. enabling the MOMIS system to exploit a multiple ontology approach with respect to the actual single ontology approach;
3. developing a new method to compute semantic mapping among source schemas in the MOMIS system.

The paper is organized as follows. Section 2, describes the proposed approach to use the Ontology-driven Semantic Mapping framework in the Global Schema generation process of MOMIS; in Section 3 future works are sketched out; finally, Section 4 is devoted to conclusions.

2 Ontology-Based Data Integration: the MOMIS-STASIS approach

This section describes our approach to use the Ontology-driven Semantic Mapping framework performed by STASIS for a different goal, i.e., during in the Global Schema Generation process performed by the MOMIS system. Intuitively, with the Ontology-driven Semantic Mapping framework we may perform in the Data Integration System the annotation of data sources elements with respect to generic ontologies (expressed in OWL), by eliminating in this way the MOMIS limitation to use only the lexical ontology WordNet. Moreover, we introduce in the MOMIS system a multiple ontology approach with respect to the actual single ontology approach. In the following, we will refer to this new approach as the MOMIS-STASIS approach.

The MOMIS-STASIS approach is shown in Figure 1. It can be divided into two macro-steps: STASIS : Semantic Link Generation (shown in Figure 1-a) and MOMIS : Global Schema Generation (shown in Figure 1-b).

2.1 STASIS : Semantic Link Generation

As stated in [2, 1] the key aspect of the STASIS framework, which distinguishes it from most existing semantic mapping approaches, is to provide an easy to use GUI, allowing users to identify semantic elements in an easy way. Once this identification has been performed STASIS lets users map their semantic entities to those of their business partners where possible assisted by STASIS. This allows users to create mappings in a more natural way by considering the meaning of elements rather than their syntactical structure. Moreover, all mappings that have been created by STASIS, as well as all semantic entities, are managed in a distributed registry and repository network. This gives STASIS another significant advantage over traditional mapping creation tools as STASIS may reuse all mappings. This allows STASIS to make some intelligent mapping suggestions by reusing mapping information from earlier semantic links.

Besides the semantic links explicitly provided by the user, an Ontology-driven Semantic Mapping approach, for the STASIS framework, has been proposed [4]. The mappings between semantic entities being used in different schemas can be achieved based on annotations linking the semantic entities with some concepts being part of an ontology. In [5], this framework has been further elaborated and it has been applied to the
Fig. 1. The MOMIS-STASIS approach for Ontology-Based Data Integration: (a) Ontology driven Semantic Mapping Discovery, (b) Global Schema Generation.

context of products and services catalogues. An overview of the process for Ontology-driven Semantic Mapping Discovery is given in Figure 1-a. It can be summed up into 3 steps (each step number is correspondingly represented in figure): (1) obtaining a neutral schema representation, (2) local source annotation, and (3) semantic mapping discovery.

Step 1. Obtaining a neutral schema representation

As sketched in Figure 1-a, the STASIS framework works on a neutral representation, which abstracts from the specific syntax and data model of a particular schema definition; therefore, all the structural and semi-structural local sources first need to be expressed in a neutral format. The neutral representation is obtained by describing the local schemas through a unified data model called Logical Data Model (LDM). For the purpose of this paper, we abstract from the specific features of LDM and we consider that this model contains common aspects of most semantic data models: it allows
the representation of *classes* (or concepts) i.e. unary predicates over individuals, *relationships* (or object properties) i.e. binary predicates relating individuals, and *attributes* (or data-type properties) i.e. binary predicates relating individuals with values such as integers and strings; classes are organized in the familiar *is-a* hierarchy. *Classes, relationships* and *attributes* are called *semantic entities*.

**Step 2. Local source annotation**

The proposed mapping process identifies mappings between semantic entities through a “reasoning” with respect to aligned ontologies. For this purpose the semantic entities need to be annotated with respect to one or more ontologies. More formally, an *annotation element* is a 4-tuple \(< ID, SE, R, concept >\) where \(ID\) is a unique identifier of the given annotation element; \(SE\) is a semantic entity of the schema; \(concept\) is a concept of the ontology; \(R\) specifies the semantic relationship which may hold between \(SE\) and \(concept\). The following semantic relationships between semantic entities and the concepts of the ontology are used: equivalence \((AR_{EQUIV})\); more general \((AR_{SUP})\); less general \((AR_{SUB})\); disjointness \((AR_{DISJ})\).

Actually within the STASIS framework are implemented only simple automatic annotation techniques, e.g. the “name-based technique” where the annotation between a semantic entity and a ontology concept is discovered by comparing only the strings of their names. The main drawback of this automatic technique is due to the existence of *synonyms* (when different words are used to name the same entities, e.g. “Last Name” and “Surname”) and *homonyms* (when the same words is used to name different entities, e.g. “peer” has a sense “equal” as well as another sense “member of nobility”) [9]. For these reason the designer has to manually refine the annotations in order to capture the semantics associated to each entities. In Section 3 a preliminary idea to overcome this limitation is described.

**Step 3. Semantic mapping discovery**

Based on the annotations made with respect to the ontologies and on the logic relationships identified between these aligned ontologies, reasoning can identify correspondences among the semantic entities and support the mapping process. Given two schemas \(S1\) and \(S2\), and assuming that OntologyA and OntologyB are the reference ontologies which have been used to annotate the content of \(S1\) and \(S2\) respectively, given a mapping between OntologyA and OntologyB which provides a correspondence between concepts and relationships in the two ontologies, a semantic mapping between the annotated schemas \(S1\) and \(S2\) is derived. The following semantic mappings between entities of two source schemas (called *semantic link* - SL) can be discovered: equivalence \((EQUIV)\); more general \((SUP)\); less general \((SUB)\); disjointness \((DISJ)\); this definition is based on the general framework proposed in [10].

More formally, an SL is a 4-tuple \(< ID, semantic\_entity1, R, semantic\_entity2 >\), where \(ID\) is a unique identifier of the given mapping element; \(semantic\_entity1\) is an entity of the first local schema; \(R\) specifies the semantic relationship which may hold between \(semantic\_entity1\) and \(semantic\_entity2\); \(semantic\_entity2\) is an entity of the second local schema.
An application example of the Ontology Driven Semantic Mapping approach is described in Section 2.3; other examples can be found in [5].

2.2 MOMIS: Global Schema Generation

The MOMIS Data Integration System, which performs information extraction and integration from both structured and semi-structured data sources, is presented in [7, 6]. Information integration is performed in a semi-automatic way, by exploiting the semantic links among source schemas and using a combination of clustering techniques and Description Logics. This integration process gives rise to a virtual integrated view of the underlying sources, where mapping rules and integrity constraints are specified to handle heterogeneity. Given a set of data sources related to a domain it is thus possible to synthesize - in a semi-automatic way - a Global Schema (called Global Virtual View - GVV) and the mappings among the local sources and the GVV. Mappings among source schemas and the GVV are defined with a Global-As-View (GAV) approach: each class of the GVV is characterized in terms of a view over the sources.

In the MOMIS System, semantic links among source schemas are mostly derived with lexicon techniques based on the lexical annotation with respect to WordNet; then, all these semantic links are collected in a Common Thesaurus.

In this paper we consider as semantic links among source schemas the semantic links defined with the STASIS framework; in other words, we consider as input of the GVV generation process the Common Thesaurus SLs generated by the STASIS framework. An overview of this GVV generation process is given in Figure 1-b.

In the GVV generation process classes describing the same or semantically related concepts in different sources are identified and clustered in the same global class. Exploiting the Common Thesaurus SLs and the local sources schemas, our approach generates a GVV consisting of a set of global classes, plus a Mapping Table (MT) for each global class, which contains the mappings to connect the global attributes of each global class with the local sources’ attributes. A MT is a table where the columns represent the local classes (LG) belonging to the global class G and whose rows represent the global attributes of G. An element $MT[GA][L]$ represents the set of local attributes of the local source L which are mapped onto the global attribute GA. The ontology designer may interactively refine and complete the proposed integration results; in particular, the mappings which has been automatically created by the system can be fine tuned in the Mapping Refinement step. Intuitively, the GVV is the intensional representation of the information provided by the Integration System, whereas the mapping assertions specify how such an intensional representation relates to the local sources managed by the Integration System.

A query posed by a user with respect to a global class is rewritten as an equivalent set of queries expressed on the local schemas (local queries); this query translation is carried out by considering the mappings between the GVV and the local schemas. Results from the local sources are then merged exploiting reconciliation techniques and proposed to the user.

For a complete description of the methodology to build and query the GVV see [7, 6].

6
2.3 Example

As a simple example let us consider two relational local sources L1 and L2, where each schema contains a relation describing purchase orders:

L1: PURCHASE_ORDER(ORDERID, BILLING_ADDRESS, DELIVERY_ADDRESS, DATE)
L2: ORDER(NUMBER, CUSTOMER_LOCATION, YEAR, MONTH, DAY)

In the following, we will described step by step the application of the MOMIS-STASIS approach on these two sources.

STASIS: Semantic Link Generation

Step 1. Obtaining a neutral schema representation
During this step the local sources L1 and L2 are translated in the neutral representation and are represented in LDM data model; for a complete and formal description of such representation see [4], where a similar example was discussed. As said before, for the purpose of this paper, we consider that the local schema L1 contains a class PURCHASE_ORDER with attributes ORDERID, BILLING_ADDRESS, DELIVERY_ADDRESS.

In this way L1.PURCHASE_ORDER, L1.PURCHASE_ORDER.BILLING_ADDRESS, L1.PURCHASE_ORDER.DELIVERY_ADDRESS etc. are semantic entities. In the same way the local schema L2 contains a class ORDER with attributes NUMBER, CUSTOMER_LOCATION, YEAR, MONTH, DAY.

Step 2. Local Source Annotation
For the sake of simplicity, we consider the annotation of schemas and the derivation of mappings with respect to a single common ontology ("Ontology-based schema mapping with a single common ontology" scenario considered in [4]). Let us give some examples of annotations of the above schemas with respect to the Purchase Order Ontology shown in Figure 2. In the examples the identifier ID is omitted and a concept C of the ontology is denoted by "O:C". In a simple annotation the concept O:C is a primitive concept or a primitive role of the ontology (e.g. the class O:ADDRESS or the property O:BILLING). In a complex annotation the concept O:C is obtained by using the OWL language constructs (e.g. "O:ADDRESS and BILLING-1.Purchase_Order" where BILLING-1 denotes the inverse of the property O:BILLING).

The following are examples of simple annotations:

(L1.PURCHASE_ORDER.BILLING_ADDRESS, AR_EQUIV, O:ADDRESS)
and
(L1.PURCHASE_ORDER.BILLING_ADDRESS, AR_EQUIV, O:BILLING).

These annotations are automatically discovered by applying the automatic "name-based" technique (see Section 2.1). However, as this technique does not consider the semantics associated to each entities, the following annotation

(L2.ORDER.CUSTOMER_LOCATION, AR_EQUIV, O:ADDRESS)
is not discovered: the entities CUSTOMER_LOCATION and the concept ADDRESS have complete different names but, in this context, they have the same senses. In Section 3 a preliminary idea to overcome this problem is described.

An example of complex annotation is

(L1.PURCHASE_ORDER.DELIVERY_ADDRESS, AR_EQUIV, O:Address and Shipping-1.Purchase_Order)

which can be considered as a refinement by the designer of the above simple annotations to state that the address in the PURCHASE_ORDER table is the “address of the Shipping in a Purchase Order”.

Other examples of complex annotations are:

(L1.PURCHASE_ORDER.BILLING_ADDRESS, AR_EQUIV, O:Address and Billing-1.Purchase_Order)

where is explicitly declared by the designer to state that the address in the PURCHASE_ORDER table is the “address of the Billing in a Purchase Order”.

(L2.ORDER.CUSTOMER_LOCATION, AR_EQUIV, O:Address and Shipping-1.Purchase_Order)

where is explicitly declared by the designer to state that the address in the ORDER table is the “address of the Shipping in a Purchase_Order”.

Fig. 2. The ontology of Purchase_order
Moreover, the designer supplies also the annotations with respect to the ontology for the semantic entities L1.PURCHASE_ORDER.ORDERID, L1.PURCHASE_ORDER.DATE and L2.ORDER.NUMBER, L2.ORDER.YEAR, L2.ORDER.MONTH, L2.ORDER.DAY.

**Step 3. Semantic mapping discovery**

From the previous annotations, for example, the following semantic link is derived:

\[(L2.ORDER.CUSTOMER_LOCATION, \text{EQUIV}, L1.PURCHASE_ORDER.DELIVERY_ADDRESS)\]

while no semantic link among CUSTOMER_LOCATION and BILLING_ADDRESS is generated.

**MOMIS: Global Schema Generation**

Given the set of semantic links described above and collected in the Common Thesaurus, the GVV is automatically generated and the classes describing the same or semantically related concepts in different sources are identified and clustered in the same global class. Moreover, the Mapping Table shown in Table 1 is automatically created by the MOMIS-STASIS approach.

<table>
<thead>
<tr>
<th>Global attributes</th>
<th>Local attributes</th>
<th>Local attributes</th>
</tr>
</thead>
<tbody>
<tr>
<td>ORDER</td>
<td>ORDER</td>
<td>PURCHASE_ORDER</td>
</tr>
<tr>
<td>NUMBER</td>
<td>NUMBER</td>
<td>ORDER_ID</td>
</tr>
<tr>
<td>DATE</td>
<td>YEAR,MONTH, DAY</td>
<td>DATE</td>
</tr>
<tr>
<td>CUSTOMER_LOCATION</td>
<td>CUSTOMER_LOCATION</td>
<td>DELIVERY_ADDRESS</td>
</tr>
<tr>
<td>BILLING_ADDRESS</td>
<td>NULL</td>
<td>BILLING_ADDRESS</td>
</tr>
</tbody>
</table>

Table 1. Mapping Table example

The global class ORDER is mapped to the local class ORDER of the L1 source and to the local class PURCHASE_ORDER of the L2 source. The NUMBER, DATE and CUSTOMER_ADDRESS global attributes are mapped to both the sources, the BILLING_ADDRESS global attribute is mapped only to the L2 source.

### 3 Future Work

One of the main advantage of the proposed approach is an accurate annotation of the schemas that produces more reliable relationships among semantic entities. The relationships among semantic entities are then exploited in order to obtain a more effective integration process. On the other hand, this more accurate annotation has the disadvantage that is currently performed manually by the integration designer.

In this work, we describe only a preliminary idea to overcome the problem of manual annotation, which will be the main subject of our future research.

Several works about automatic annotation are proposed in literature but only a few of them are applied in the context of schemas/ontologies matching discovery. In [13,
where we introduced a mapping technique based on Lexical Knowledge Extraction: first, an Automatic Lexical Annotation method is applied to annotate, with respect to WordNet, schemas/ontologies elements then lexical relationships are extracted based on such annotations.

Moreover, in [15] a way to apply our Automatic Lexical Annotation method to the SCARLET matcher [3], is presented. SCARLET is a technique for discovering relationships between two concepts by making use of online available ontologies. The matcher can discover semantic relationships by reusing knowledge declared within a single ontology or by combining knowledge contained in several ontologies. By applying Automatic Lexical Annotation based on WordNet, the matcher validates the discovered mapping by exploring the semantics of the terms involved in the matching process.

Starting from these works, we agree that the WordNet lexical ontology can be used to improve the annotation phase of the Ontology Driven Semantic Mapping process. The strength of an lexical ontology like WordNet is the presence of a wide network of semantic relationships among the different words meanings, which represent a key element for automatic annotation techniques.

Let us consider the example shows in Figure 3: during the local source annotation step, the relationships between semantic entities and ontology concepts have to be discovered. Our idea can be summed up in three main steps (each step number is correspondingly represented in figure):

1. **Ontology and local source annotation with respect to WordNet**: both the ontology and the local source, are annotated, with respect to WordNet, by using the Automatic Lexical Annotation method described in [13]: e.g., as shown in Figure 3, the semantic entity “Surname” is annotated with the first sense in WordNet (indicated in Figure with “#1”) for the word “SURNAME” and the ontology concept “LastName” is annotated with the first sense in WordNet for the word “LASTNAME”;

2. **WordNet semantic relationship discovery**: starting from the previous annotations, a set of WordNet semantic relationships (synonym (equivalence), hypernym (more general) etc.) is discovered among semantic entities and ontology concepts: e.g., as shown in Figure 3, a synonym relationship is discovered between the semantic entity “Surname” and the ontology concept “LastName”.

3. **Local source annotation for Ontology Driven Semantic Mapping**: starting from the set of WordNet semantic relationships previously discovered, a correspondent set of annotation for Ontology-Driven Semantic Mapping can be discovered: e.g., starting from the WordNet synonym relationship between “Surname” and the “LastName”, the following annotation is established (the annotation unique identifier ID is omitted):

   (surname, AR_EQUIV, O:LastName)

In this way, we can automatically annotate a set of local schema semantic entities with respect to the considered ontology. However, these annotations can be incomplete (because WordNet does not contain many domain dependent words) and require a designer refinement. Even if this preliminary idea needs to be further investigated, it represents a fundamental start point to help the designer during the time consuming task of manual annotation.
Another future work will be the investigation of automatic techniques to discover the relationships among semantic entities combining the exploration of multiple and heterogeneous online ontologies with the annotations provided by the WordNet lexical ontology. The use of online ontologies represents an effective way to improve the semantic mapping process. For example, in [16], automatic techniques to discover the relationships between two concepts automatically finding and exploring multiple and heterogeneous online ontologies, have been proposed.

Fig. 3. A preliminary idea to perform automatic annotation for Ontology-driven Semantic Mapping

4 Conclusions

In this paper, we have described the early effort to obtain an effective approach for Ontology-Based Data Integration combining the techniques provided by the MOMIS and the STASIS frameworks. In particular, with the Ontology-driven Semantic Mapping framework we have performed in the Data Integration System the annotation of data sources elements with respect to generic ontologies (expressed in OWL), by eliminating in this way the MOMIS limitation to use only the lexical ontology WordNet. Moreover, we have introduced in the MOMIS system a multiple ontology approach with respect to the actual single ontology approach. Even if this work needs to be further investigated (as described in Section 3), it represents a fundamental start point versus an automatic Ontology-Based Data Integration System.

5 Acknowledgment

This work supported by STASIS IST project (FP6-2005-IST-5-034980).
References


