An Ontology-Based Data Integration system for data and multimedia sources

Domenico Beneventano, Mirko Orsini, Laura Po, Antonio Sala, Serena Sorrentino
DII, University of Modena and Reggio Emilia
via Vignolese 905, 41125 Modena, Italy
Email: firstname.lastname@unimore.it

Abstract—**Data integration is the problem of combining data residing at distributed heterogeneous sources, including multimedia sources, and providing the user with a unified view of these data. Ontology based Data Integration involves the use of ontology(s) to effectively combine data and information from multiple heterogeneous sources [16]. Ontologies, with respect to the integration of data sources, can be used for the identification and association of semantically corresponding information concepts, i.e. for the definition of semantic mappings among concepts of the information sources. MOMIS is a Data Integration System which performs in-formation extraction and integration from both structured and semi-structured data sources [6]. In [5] MOMIS was extended to manage "traditional" and "multimedia" data sources at the same time. STASIS is 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 pro-posed [7]. 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 for data and multimedia sources.**

Keywords-**data integration; ontology; semantic mappings; multimedia data;**

I. Introduction

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 [17] based on a set of data sources and a global schema (Global Virtual View-GVV) which provides a reconciled, integrated, and virtual view of the underlying sources; modeling the mappings among sources and the GVV is a crucial aspect.

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 [6], [4]. The integration process gives rise to a GVV for which mapping rules and integrity constraints are specified to handle heterogeneity. In [8], [5] MOMIS has been extended to manage "traditional" and "multimedia" data sources at the same time; the result has been implemented in a tool for integrating traditional and multimedia data sources in a GVV which can be transparently queried by users. We believe this is an interesting achievement for several reasons. Firstly, the application domain: there are several use cases where joining traditional and multimedia data is relevant. Secondly, multimedia and traditional data sources are usually represented with different models. While there is a rich literature for transforming the differently modelled traditional data sources into a common model and it is possible to represent different multimedia sources with a uniform standard model such as MPEG-7, a standard for representing traditional and multimedia data does not exist. Finally, while different languages and different interfaces for querying "traditional" and "multimedia" data sources have been developed, a framework for querying either traditional and multimedia data does not exist.

Ontologies can be used in an integration task to describe the semantics of the information sources and to make the contents explicit [16]. With respect to the integration of data sources, they can be used for the identification and association of semantically corresponding information concepts.

In [16], 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 [16].

With respect to the above classification, the MOMIS Data Integration System uses a single ontology approach, where the lexical ontology WordNet [13] 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.

In this paper, mainly to overcome this limitation, we propose to combine the MOMIS framework with the STASIS framework.

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 [7].

In [9] 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 Global Schema Generation approach for Ontology-Based Data Integration for data and multimedia sources. 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 II, describes the proposed approach to use the Ontology-driven Semantic Mapping framework in the Global Schema generation process of MOMIS and section III is devoted to conclusions and future work.

II. 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. 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).

A. 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.

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. 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*. For multimedia source this neutral description is obtained by means of the MOMIS/MILOS system, as described in section II-B1.

Step 2. Local source annotation
The proposed mapping process identifies mappings between semantic entities through a "reasoning" with respect to aligned ontologies. Semantics of the data is captured by some kind of *semantic correspondences* between the database schema and ontologies. For this purpose the

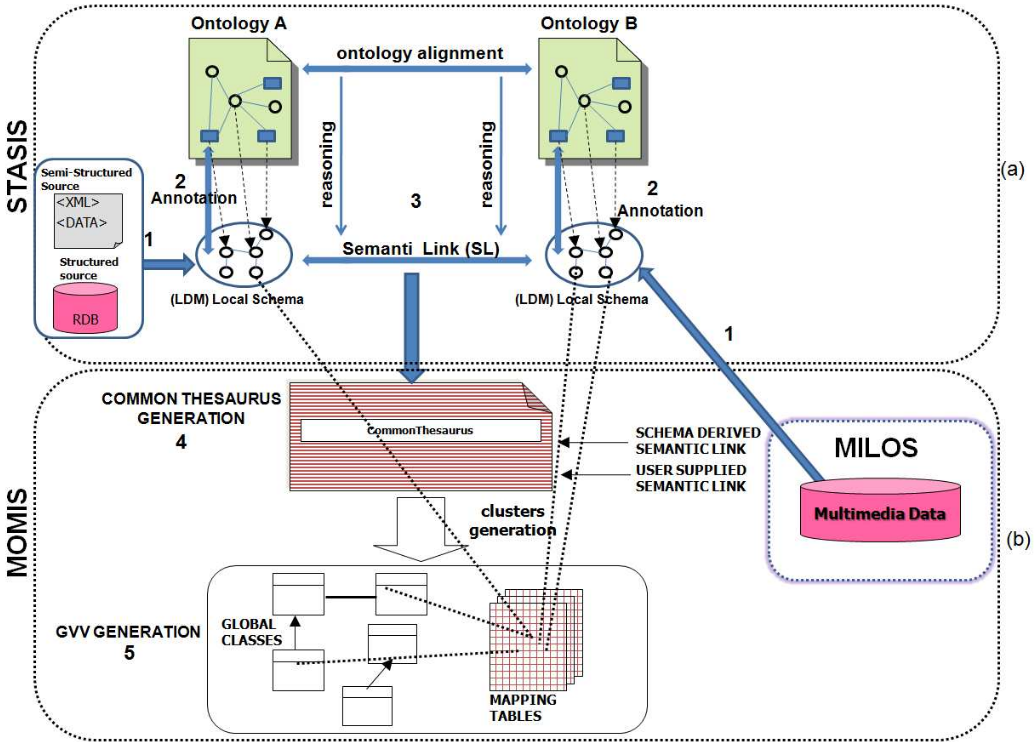

Figure 1. The MOMIS-STASIS approach for Ontology-Based Data Integration: (a) Ontology driven Semantic Mapping Discovery, (b) Global Schema Generation.

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 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. For these reason, the designer have to manually refine the annotations in order to capture the semantics associated to each entities.

Step 3. Semantic mapping discovery

Based on the annotation 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 II-C; other examples can be found in [9].

B. MOMIS: Global Schema Generation

In the MOMIS Data Integration System, information integration is performed by exploiting the semantic links among source schemas and using clustering techniques. Given a set of data sources 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 source schemas and the GVV [6], [4]. Source schemas and the GVV are described in ODL$_{I^3}$ which is very close to the standard ODL language [1] and shares with the LDM model of STASIS the basic features, such as classes, relationships and attributes. As a consequence, the translation from ODL$_{I^3}$ to LDM (and viceversa) is straightforward.

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 in Figure 1-b.

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 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. An example of this process will be shown in next section. The integration 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.

MOMIS follows a Global-As-View (GAV) approach [11], [15], then the GVV is designed to be a view over the local sources: each class of the GVV is characterized in terms of a view over its local classes. On the basis of this view, a query posed by a user with respect to the global class can be rewritten as an equivalent set of queries (local queries) expressed on the local classes. The local query answers are then merged exploiting reconciliation techniques and proposed to the user.

The definition of the view associated to a global class and the related querying problem are out of the scope of this paper; for a complete description of the methodology to build and query the GVV see [6], [4].

1) MOMIS extension for multimedia data sources: In [8], [5] MOMIS has been extended to manage "traditional" and "multimedia" data sources at the same time. The extension is based on MILOS systems [3], for managing the interaction with the multimedia sources. MILOS is a Multimedia Content Management System tailored to support design and effective implementation of digital library applications; MILOS supports the storage and content based retrieval of any multimedia documents whose descriptions are provided by using arbitrary metadata models represented in XML. In particular, in [8], [5] a notion of DMS (*Data and Multimedia Source*) is introduced to represent and query data source and multimedia sources in a uniform way. A DMS is represented with a local schema and in ODL$_{I^3}$, each class of a DMS schema, in general, includes a set of attributes declared using standard predefined types (such as string, double, integer, etc.). Along with these *standard attributes*, a DMS includes *multimedia attributes*, declared by means of special predefined classes in ODL$_{I^3}$, which support similarity based searches.

In both the steps of Semantic Link Generation and the GVV generation described before multimedia sources are treated the same way as data sources; in particular, standard (multimedia) attributes can be mapped only into standard (multimedia) attributes. Then, in the brief example of next section, we do not consider multimedia sources. The differences between data sources and multimedia sources, and between standard and multimedia attributes, is relevant in the querying phase (described in [8], [5]) which is out of the scope of this paper.

C. 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 MOMIS-STASIS Global Schema Generation approach on these two local sources.

STASIS: Semantic Link Generation

Step 1. 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 a such representation see [7], 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, Date`.

In this way `L1.Purchase_Order, L1.Billing_Address, L1.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

[1] http://www.odmg.org

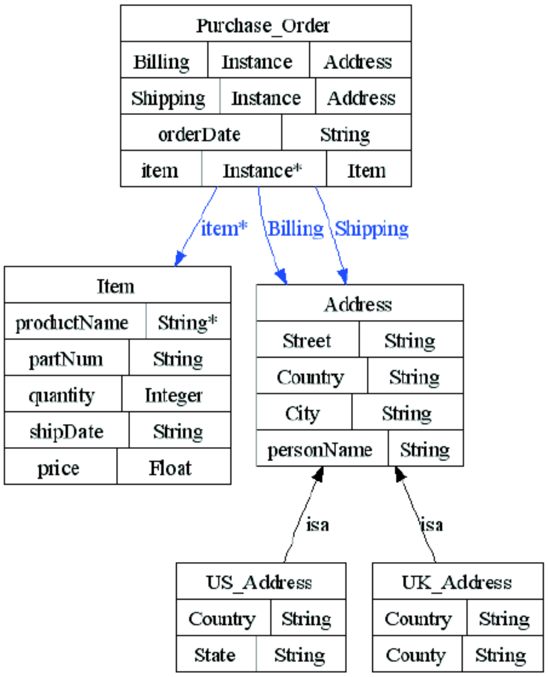

Figure 2. The ontology of Purchase_order

single common ontology ("Ontology-based schema mapping with a single common ontology" scenario considered in [7]).

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.Billing_Address, AR_EQUIV, O:Address)
```

and

```
(L1.Billing_Address, AR_EQUIV, O:Billing).
```

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

```
(L2.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 III a preliminary idea to overcome this problem is described.

An example of complex annotation is

```
(L1.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.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.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".

Moreover, the designer supplies also the annotations with respect to the ontology for the semantic entities L1.OrderID, L1.Date and L2.Number, L2.Year, L2.Month, L2.Day.

Step 3. Semantic mapping discovery
From the previous annotations, for example, the following semantic link is derived:

```
(L2.Customer_Location, EQUIV,
 L1.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 clusterized in the same global class. Moreover, the Mapping Table shown in Table I is automatically created by the MOMIS-STASIS approach. 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.

III. Conclusions and Future Work

In this paper, we have described the early effort to obtain an effective Global Schema Generation approach for Ontology-Based Data Integration for data and multimedia sources, combining the techniques provided by the MOMIS and the STASIS frameworks. In particular, with

Global attributes *ORDER*	Local attributes *ORDER*	Local attributes *PURCHASE_ORDER*
NUMBER	NUMBER	ORDER_ID
DATE	YEAR,MONTH,DAY	DATE
CUSTOMER_LOCATION	CUSTOMER_LOCATION	DELIVERY_ADDRESS
BILLING_ADDRESS	*NULL*	BILLING_ADDRESS

Table I
MAPPING TABLE EXAMPLE

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). In this way, we have eliminated the MOMIS limitation to use only the lexical ontology WordNet by introducing a *multiple ontology* approach with respect to the actual *single ontology* approach.

One of the main advantage of the proposed approach is an accurate annotation of the schemas that produces more reliable relationships among semantic entities. On the other hand, this more accurate annotation has the disadvantage that is essentially performed manually by the integration designer. For this reason, future work will be devoted to improve the annotation phase by studying automatic lexical annotation techniques. 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 [14].

ACKNOWLEDGMENT

This work has been supported by STASIS IST project (FP6-2005-IST-5-034980) http://www.stasis-project.net.

REFERENCES

[1] Sven Abels, Stuart Campbell, and Hamzeh Sheikhhasan. Stasis - creating an eclipse based semantic mapping platform. In *eChallenges 2008*.

[2] Sven Abels, Hamzeh Sheikhhasan, and Paul Cranner. Simplifying e-business collaboration by providing a semantic mapping platform. In *I-ESA '08 Workshop*, 2008.

[3] Giuseppe Amato, Claudio Gennaro, Pasquale Savino, and Fausto Rabitti. Milos: a Multimedia Content Management System for Digital Library Applications. In *Proceedings of ECDL 2004*.

[4] Domenico Beneventano and Sonia Bergamaschi. Semantic search engines based on data integration systems. In *Semantic Web Services: Theory, Tools and Applications*. Idea Group Publishing, 2006.

[5] Domenico Beneventano, Sonia Bergamaschi, Claudio Gennaro, Francesco Guerra, Matteo Mordacchini, and Antonio Sala. A mediator system for data and multimedia sources. In *Workshop on Data Integration through Semantic Technology at the 3rd Asian Semantic Web Conference*, 2008.

[6] Domenico Beneventano, Sonia Bergamaschi, Francesco Guerra, and Maurizio Vincini. Synthesizing an integrated ontology. *IEEE Internet Computing*, 7(5):42–51, 2003.

[7] Domenico Beneventano, Nikolai Dahlem, Sabina El Haoum, Alex Hahn, Daniele Montanari, and Matthias Reinelt. Ontology-driven semantic mapping. In *I-ESA 2008*, pages 329–342, 2008.

[8] Domenico Beneventano, Claudio Gennaro, and Francesco Guerra. A methodology for building and querying an ontology representing data and multimedia sources. In *VLDB-Workshop ODBIS 2008.*

[9] Domenico Beneventano and Daniele Montanari. Ontological mappings of product catalogues. In *Ontology Matching ISWC-Workshop (OM 2008)*, pages 244–249, 2008.

[10] Fausto Giunchiglia, Mikalai Yatskevich, and Pavel Shvaiko. Semantic matching: Algorithms and implementation. *J. Data Semantics*, 9:1–38, 2007.

[11] Alon Y. Halevy. Answering queries using views: A survey. *VLDB Journal*, 10(4):270–294, 2001.

[12] Maurizio Lenzerini. Data integration: A theoretical perspective. In Lucian Popa, editor, *PODS*, pages 233–246. ACM, 2002.

[13] George A. Miller. WordNet: A Lexical Database for English. *Communications of the ACM*, 38(11):39–41, 1995.

[14] Marta Sabou, Mathieu Daquin, and Enrico Motta. Exploring the semantic web as background knowledge for ontology matching. *Journal on Data Semantics XI*, pages 156–190, 2008.

[15] Jeffrey D. Ullman. Information integration using logical views. *Theor. Comput. Sci.*, 239(2):189–210, 2000.

[16] H. Wache, T. Vgele, . Visser, H. Stuckenschmidt, G. Schuster, H. Neumann, and S. Hbner. Ontology-based integration of information - a survey of existing approaches. pages 108–117, 2001.

[17] Gio Wiederhold. Mediators in the architecture of future information systems. *IEEE Computer*, 25(3):38–49, 1992.