A Semantic Approach to Information Integration: the MOMIS project*

S. Bergamaschi[1], S. Castano[2], S. De Capitani di Vimercati[2],
S. Montanari[1], M. Vincini[1]

(1) University of Modena
DSI - Via Campi 213/B - 41100 Modena
e-mail: [sonia,montanar,vincini]@dsi.unimo.it

(2) University of Milano
DSI - Via Comelico, 39 - 20135 Milano
e-mail: [castano,decapita]@dsi.unimi.it

1 Introduction

In the continuing quest to provide uniform access to distributed information, the problem of integrating information from heterogeneous sources is becoming more and more relevant.

Main problems to be faced in integrating information coming from distributed sources are related to structural and implementation heterogeneity (including differences in hardware platforms, DBMS, data models and data languages), and to the lack of a common ontology, which leads to semantic heterogeneity. Semantic heterogeneity occurs when different names are employed to represent the same information or when different modeling constructs are used to represent the same piece of information in different sources [15].
Some approaches have been recently proposed in the literature for the extraction and integration of conventional structured databases [7] and semi-structured data [8, 12] taking into account semantic heterogeneity. Data integration architectures are usually based on *mediators* [17], where knowledge about data of multiple sources is combined to provide a global view of the underlying data. Two fundamental approaches have emerged in the literature: *structural* [14, 16] and *semantic* [1, 3, 7].

In this paper, we propose a semantic approach to the integration of heterogeneous information. The approach follows the *semantic* paradigm, in that conceptual schemata of an involved source are considered, and a common data model (ODM_{I^3}) and language (ODL_{I^3}) are adopted to describe sharable information. ODM_{I^3} and ODL_{I^3} are defined as a subset of the corresponding ODMG-93 [13] ODM and ODL. A Description Logics `ocdl` (object description language with constraints [6]) is used as a kernel language and ODB-Tools as the supporting system [2].

We developed an I^3 system, called MOMIS (**M**ediator envir**O**nment for **M**ultiple **I**nformation **S**ources) implementing our approach [4].

Information integration in MOMIS is based on schemata and is performed through an *extraction and analysis* process followed by a *unification* process. The extraction and analysis process is devoted to derive a Common Thesaurus of *terminological relationships*, based on ODL_{I^3} schemata descriptions, and to the construction of clusters (by means of clustering techniques) of ODL_{I^3} classes, describing similar information in different schemata. The unification process builds an integrated global schema for the analyzed sources by integrating ODL_{I^3} classes in a given cluster.

The use of the `ocdl` Description Logics together with hierarchical clustering techniques are the original contributions of the approach to enhance a semi-automated integration process. Description Logics allows us to interactively set-up the Thesaurus by deriving explicit terminological relationships from ODL_{I^3} schemata descriptions and by inferring new relationships out of the explicit ones. Moreover, optimization of the queries against the global schema is possible using description logics. Clustering techniques allow the automated identification of ODL_{I^3} classes in different source schemata that are semantically related and thus candidate to be unified in the global schema.

2 The MOMIS architecture and the ODL_{I^3} language

In Fig. 1 the architecture of the MOMIS system is shown. With respect to the literature, this can be considered as an example of powerful I^3 sys-

*This research has been partially funded by the INTERDATA - MURST 40% 97-98 project.

tem [9] and follows the TSIMMIS architecture [14]. Above each source lies a translator (called *wrapper*) responsible for translating the structure of the data source into the common ODL$_{I^3}$ language. In a similar way, the *wrapper* performs a translation of the query from the OQL$_{I^3}$ language to a local request to be executed by a single source. Above the *wrapper* there is a *mediator*, a software module that combines, integrates, and refines ODL$_{I^3}$ schemata received from the *wrappers*. In addition, the mediator generates the OQL$_{I^3}$ queries for the *wrappers*, starting from the query request formulated with respect to the global schema. The mediator module is obtained by coupling a *semantic approach*, based on a Description Logics component, i.e. ODB-Tools Engine [2] (available at http://www.sparc20.dsi.unimo.it/publications.html), and an extraction and analysis component, i.e., Schema Analyzer and Classifier, together with a minimal ODL$_{I^3}$ interface.

In order to easily communicate source descriptions between wrappers and mediator engine, we introduced a data description language, called ODL$_{I^3}$.

According to the recommendations of [9], and to the diffusion of the object data model (and its standard ODMG-93), ODL$_{I^3}$ is very close to the ODL language, supporting requirements of our intelligent information integration system. ODL$_{I^3}$ is a source independent language used by the mediator to manage the system in a common way (we suppose to deal with different source types, such as relational databases, object-oriented databases, files). The main extension, w.r.t. ODL, is the capability of expressing two kind of rules: *if then* rules, expressing in a declarative way integrity constraints intra and inter sources, and *mapping* rules between sources. It will be the wrapper task to translate the data description language of any particular source into ODL$_{I^3}$ description, and to add information needed by the mediator, such as the source name and type.

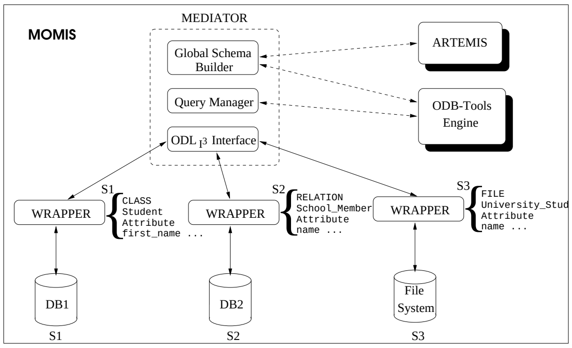

Figure 1: Architecture of the MOMIS I^3 system

3 Overview of the approach

The MOMIS approach to intelligent schema integration is articulated in the following phases:

1. *Generation of a Common Thesaurus.*
   The objective of this step is the construction of a Common Thesaurus of *terminological relationships* for schema classes belonging to different source ODL$_{I^3}$ schemata.
   The following kinds of terminological relationships are specified: SYN (Synonym-of), defined between two terms t_i and t_j, with $t_i \neq t_j$, that are considered synonyms, i.e., that can be interchangeably used in every considered source, without changes in meaning; BT (Broader Terms), or hypernymy, defined between two terms t_i and t_j such as t_i has a broader, more general meaning than t_j; RT (Related Terms), or positive association, defined between two terms t_i and t_j that are generally used together in the same context.
   Terminological relationships are derived in a semi-automatic way, by analyzing the structure and context of classes in the schema, by using ODB-Tools and the Description Logics techniques.

2. *Affinity analysis of ODL$_{I^3}$ classes.*
   Terminological relationships in the Thesaurus are used to evaluate the level of *affinity* between classes intra and inter sources. The concept of affinity is introduced to formalize the kind of relationships that can occur between classes from the integration point of view. The affinity of two classes is established by means of affinity coefficients based on class names and attributes [10, 11].

3. *Clustering ODL$_{I^3}$ classes .*
   Classes with affinity in different sources are grouped together in clusters using hierarchical clustering techniques. The goal is to identify the classes that have to be integrated since describing the same or semantically related information.

4. *Generation of the mediator global schema.*
   Unification of affinity clusters leads to the construction of the global schema of the mediator. A class is defined for each cluster, which

is representative of all cluster's classes and is characterized by the union of their attributes. The global schema for the analyzed sources is composed of all these new classes, derived from clusters, and is the basis for posing queries against the sources.

Each phase of the integration process is described in detail in [4, 5]. Once the mediator global schema has been constructed, it can be exploited by the users for posing queries. The information on the global schema is used by the *Query Manager* module of MOMIS for query reformulation and for semantic optimization using ODB-Tools, as discussed in [4].

4 Generation of the mediator global schema

In this section we present the process which leads to the definition of the mediator global schema, that is the mediator view of data stored in local sources.

4.1 Running example

First, we introduce an example used in the paper to explain the approach (see Fig. 2). We consider three different sources. The first source is a relational database, `University` (S_1), containing information about the staff and the students of a given university. The second source `Computer_Science` (S_2) contains information about people belonging to the computer science department of the same university, and is an object-oriented database. A third source is also available, `Tax_Position` (S_3), derived from the registrar's office. It consists of a file system, storing information about student's tax_fees. For the complete source descriptions see [4].

The generation of Common Thesaurus, the Affinity analysis and the Clustering techniques applied to our example give raise to the cluster tree shown in Fig. 3, which classifies classes into groups at different level of affinity.

Starting from the affinity tree produced with clustering, we define, for each cluster in the tree, a global class *global_class$_i$* representative of the classes contained in the cluster (i.e., a class providing the unified view of all the classes of the cluster).

The generation of the *global_class$_i$* is interactive with the designer. Let Cl_i be a cluster in the affinity tree. First, the Global Schema Builder component of MOMIS associates to the *global_class$_i$* a set of global attributes, corresponding to the union of

University source ($\mathbf{S_1}$)

```
Research_Staff(first_name,last_name,relation,
email,dept_code,section_code)
School_Member(first_name,last_name,faculty,year)
Department(dept_name,dept_code,budget,dept_area)
Section(section_name,section_code,length,room_code)
Room(room_code,seats_number,notes)
```

Computer_Science source ($\mathbf{S_2}$)

```
CS_Person(name)
Professor:CS_Person(title,belongs_to:Division,rank)
Student:CS_Person(year,takes:set⟨Course⟩,rank)
Division(description,address:Location,fund,
sector,employee_nr)
Location(city,street,number,county)
Course(course_name,taught_by:Professor)
```

Tax_Position source ($\mathbf{S_3}$)

```
University_Student(name,student_code,faculty_name,
tax_fee)
```

Figure 2: Example with three source schemata

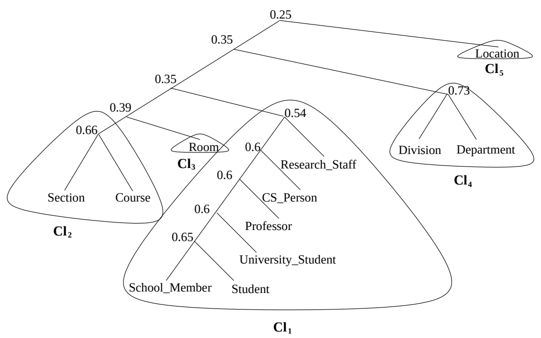

Figure 3: Affinity tree of S_1, S_2, and S_3

the attributes of the classes belonging to Cl_i, where the attributes with affinity are unified into a unique global attribute in *global_class$_i$*. The attribute unification process is performed automatically for what concerns the names of attributes with affinity, according to the following rules:

- for attributes that have name affinity due to SYN relationships, only one term is selected and assigned to the corresponding global attribute in *global_class$_i$*;
- for attributes that have name affinity due to BT and NT relationships, a name which is a broader term for all of them is selected and assigned to the corresponding global attribute in *global_class$_i$*.

For example, the output of this attribute unification process for cluster Cl_1 of Fig. 3, is the following set of global attributes:

```
Cl1 = (name, rank, title, dept_code, year,
       takes, relation,email, student_code,
       tax_fee, section_code, faculty )
```

Additional information has to be provided by the designer for completing the global class definition.

A global class *global_class*$_i$ is specified in ODL$_{I^3}$ where the information on attribute mappings and default values is represented in form of *rules*. An example of ODL$_{I^3}$ specification for the global class `University_Person` is shown in Fig. 4. As we can see from this figure, for each attribute, in addition to its declaration, mapping rules are defined, specifying both information on how to map the attribute on the corresponding attributes of the associated cluster and on possible default/null values defined for it on cluster classes. For example, for the global attribute `name`, the mapping rule specifies the attributes that have to be considered in each class of the cluster Cl_1. In this case, an *and* correspondence is defined for `name` for the class `University.Research_Staff` (we use dot notation for specifying the source of a given class belonging to the cluster). A mapping rule is defined for the global attribute `rank` to specify the value to be associated with `rank` for the instances of `University.Research_Staff` and `University.School_Member`.

The global schema of the mediator is composed of the global classes defined for all the clusters of the affinity tree.

5 Implementation of MOMIS: the Global Schema Builder

In this section, a brief description of the Global Schema Builder module is given, to outline the state of implementation of the MOMIS project. The Global Schema Builder processes the local schemata descriptions, received from the information sources, to obtain the Mediator Global Schema, that will be the base for the user's queries. It is composed by the following components (shown in Fig. 5):

1. **SIM**$_1$(**S**chemata **I**ntegrator **M**odule, first version): it reads the local schemata descriptions, expressed in the ODL$_{I^3}$ language, to derive the Common Thesaurus. In particular, a set of terminological relationships is stored in the thesaurus, by interacting with the user and by using *Description Logics* (supported by ODB-Tools) to express the logical links existing intra and inter sources;

2. **Artemis**: starting from the relationships of the Common Thesaurus, Affinity Coefficients are computed between all the pairs of local classes to be integrated, to evaluate their level of similarity. Similar classes are grouped together using *clustering tecniques*: every generated cluster will correspond to a mediator global class [10][1].

```
interface University_Person
(extent Research_Staffers, School_Members, CS_Persons
        Professors, Students, University_Students
 key    name)
{ attribute string name
   mapping_rule (University.Research_Staff.first_name and
                 University.Research_Staff.last_name),
                 (University.School_Member.first_name and
                 University.School_Member.last_name),
                 Computer_Science.CS_Person.name,
                 Computer_Science.Professor.name,
                 Computer_Science.Student.name,
                 Tax_Position.University_Student.name;
  attribute string rank
   mapping_rule University.Research_Staff = 'Professor',
                 University.School_Member = 'Student',
   ... }
```

Figure 4: Example of global class specification in ODL$_{I^3}$

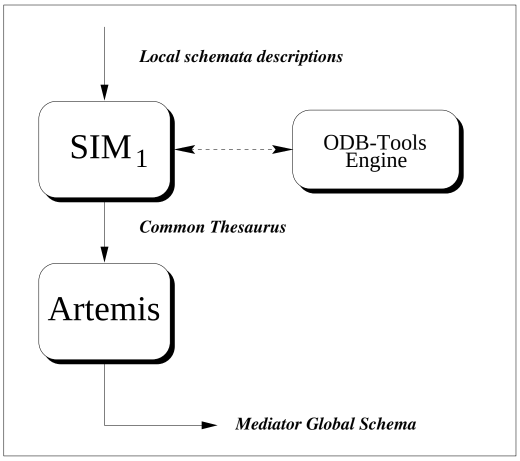

Figure 5: Global Schema Builder

6 Conclusions and future work

In this paper, we have presented an intelligent approach to schema integration for heterogeneous information sources. It is a semantic approach based on a DescriptionLogics component (ODB-Tools engine) and a cluster generator module (ARTEMIS) together with a minimal ODL$_{I^3}$ interface module.

[1]This component has been developed at the University of Milano

Future research work will be devoted to a deeper study of query processing and optimization processes, and to the development of corresponding Query Manager functionalities. In particular, interschema extensional properties relevant to perform the information integration process in view of query processing will be investigated. Moreover, the impact of this topic on clustering will be studied in order to enhance the classification process with semantic features.

References

[1] Y. Arens, C. A. Knoblock and C. Hsu, "Query Processing in the SIMS Information Mediator," in *Advanced Planning Technology, editor, Austin Tate, AAAI Press, CA*, 1996.

[2] D. Beneventano, S. Bergamaschi, C. Sartori, M. Vincini, "ODB-QOptimizer: a Tool for Semantic Query Optimization in OODB", in *Proc. of Fifth Conference of the Italian Association for Artificial Intelligence (AI*IA97)*, Rome 1997.

[3] S. Bergamaschi "Extraction of Informations from highly Heterogeneous Sources of Textual Data," in *Cooperative Information Agents, First International Workshop, CIA' 97 Proceedings. Lecture Notes in Computer Science*, Kiel, Germany, February 26-28, 1997.

[4] S.Bergamaschi, S.Castano, S.De Capitani di Vimercati, S.Montanari, and M.Vincini. An intelligent approach to information integration. *Formal Ontology in Information Systems FOIS98, june 1998.*

[5] S.Bergamaschi, S.Castano, S.De Capitani di Vimercati, S.Montanari, and M.Vincini. Exploiting schema knowledge for the integration of heterogeneous sources. *Proc. of Convegno Nazionale Sistemi Evoluti per Basi di Dati SEBD98, june 1998.*

[6] S. Bergamaschi and B. Nebel, "Acquisition and Validation of Complex Object Database Schemata Supporting Multiple Inheritance," *Applied Intelligence: The International Journal of Artificial Intelligence, Neural Networks and Complex Problem Solving Technologies*, 4:185–203, 1994.

[7] M.W. Bright, A.R. Hurson, S. Pakzad, "Automated Resolution of Semantic Heterogeneity in Multidatabases," *ACM Transactions on Database Systems,* Vol.19, No.2, June 1994, pp.212-253.

[8] P. Buneman, S. Davidson, M. Fernandez and D. Suciu, " Adding a Structure to Unstructured Data," in *Proc. of Int. Conf. on Database Theory*, Delphi, Greece, January 1997, pp. 336-350.

[9] P. Buneman, L. Raschid, J. Ullman, "Mediator Languages - a Proposal for a Standard," Report of an I^3/POB working group held at the University of Maryland, April 1996.

[10] S. Castano, V. De Antonellis, "Semantic Dictionary Design for Database Interoperability," in *Proc. of Int. Conf. on Data Engineering, ICDE'97*, Birmingham, UK, April 1997.

[11] S. Castano, V. De Antonellis, M.G. Fugini, B. Pernici, "Conceptual Schema Analysis: Techniques and Applications," accepted for publication on *ACM Transactions on Database Systems*, 1997.

[12] T. Catarci, M. Lenzerini, "Representing and using interschema knowledge in cooperative information systems", Journal of Intelligent and Cooperative Information Systems, 2(4), 375-398, 1993.

[13] R. Cattel (ed.), *The Object Database Standard: ODMG-93*, Morgan Kaufmann, 1996.

[14] J. Hammer, H. Garcia-Molina, K. Ireland, Y. Papakonstantinou, J. Ullman, and J. Widom. "Information Translation, Mediation, and Mosaic-Based Browsing in the TSIMMIS System". In Exhibits Program of the Proceedings of the ACM SIGMOD, page 483, CA 1995.

[15] W. Kim, I. Choi, S. Gala, M. Scheevel, "On Resolving Schematic Heterogeneity in Multidatabase Systems," *Distributed and Parallel Databases,* Vol.1, No.3, 1993.

[16] M.T. Roth, P. Scharz, "Don't Scrap It, Wrap it! A Wrapper Architecture for Legacy Data Sources", in *Proc. of the 23rd Int. Conf. on Very Large Databases*, Athens, Greece, 1997.

[17] G. Wiederhold, "Mediators in the architecture of Future Information Systems," *IEEE Computer*, Vol. 25, 1992, pp.38-49.