The MOMIS approach to Information Integration

1 Introduction

The web explosion, both at internet and intranet level, has transformed the electronic information system from single isolated node to an entry points into a worldwide network of information exchange and business transactions. Business and commerce has taken the opportunity of the new technologies to define the e-commerce activity. An electronic marketplace represents a virtual place where buyers and sellers meet to exchange goods and services, by sharing information that is often obtained as hypertext catalogs from different companies. Companies have equipped themselves with data storing systems building up informative systems containing data which are related one another, but which are often redundant, heterogeneous and not always substantial. The problems that have to be faced in this field are mainly due to both structural and application heterogeneity, as well as to the lack of a common ontology, causing semantic differences between information sources. Moreover, these semantic differences can cause different kinds of conflicts, ranging from simple contradictions in names' use (when different names are used by different source to indicate the same concept), to structural conflicts (when different models/primitives are used to represent the same information).

Therefore one of the main challenges for the designers of the e-commerce infrastructures is the information sharing, retrieving data located in different sources thus obtaining an integrated view to overcome any contradiction or redundancy. Virtual Catalogs synthesize this approach as they are conceived as instruments to dynamically retrieve information from multiple catalogs and present product data in a unified manner, without directly storing product data from catalogs. Customers, instead of having to interact with multiple heterogeneous catalogs, can interact in a uniform way with a virtual catalog.

In this paper we propose a designer support tool, called SI–Designer, for information integration developed within the MOMIS project. The MOMIS project (Mediator environment for Multiple Information Sources) [6, 3, 5] aims to integrate data from structured and semistructured data sources. SI–Designer is a support tool for semiautomatic integration of heterogeneous sources schema (relational, object, XML and semistructured sources); it carries out integration following a semantic approach which uses intelligent OLCD Description logics-based techniques, clustering techniques and an ODM-ODMG extended model to represent extracted and integrated information, ODM_{I^3}. Using the ODL_{I^3} language, referred to the ODM_{I^3} model, it is possible to describe the sources (local schema) and SI–Designer supports the designer in the creation of an integrated view of all the sources (Global Virtual View), which is expressed using XML standard. XML is a data-description language that meets the need to exchange data between business processes and applications without regard to source or destination platform issues. The use of XML in the definition of the Global Virtual View lets to use Momis infrastructure with other open integration information systems by the interchange of XML data files.

2 System Architecture

Like other integration projects [1, 13], MOMIS follows a "semantic approach" to information integration based on the conceptual schema, or metadata, of the information sources, and on the the I^3 architecture [11] (see figure 1). The system is composed by the following functional elements that communicates using the CORBA [10] standard:

1. a common data model, ODM_{I^3}, which is defined according to the ODL_{I^3} language, to describe source schemas for integration purposes. ODM_{I^3} and ODL_{I^3} have been defined in MOMIS as subset of the corresponding ones in ODMG, following the proposal for a standard mediator language developed by the I^3/POB working group [7]. In addition, ODL_{I^3} introduces new constructors to support the semantic integration process;

2. *Wrappers*, placed over each sources, translate metadata descriptions of the sources into the common ODL_{I^3} representation, translate (reformulate) a global query expressed in the OQL_{I^3}[1] query language into queries expressed in the sources languages and export query result data set;

3. a *Mediator*, which is composed of two modules: the *SI-Designer* and the *Query Manager* (QM). The SI-Designer module processes and integrates ODL_{I^3} descriptions received from wrappers to derive the integrated representation of the information sources. The QM module performs query processing and optimization. The QM generates OQL_{I^3} queries to be sent to

[1] OQL_{I^3} is a subset of OQL-ODMG.

wrappers starting from each query posed by the user on the Global Schema. QM automatically generates the translation of the query into a corresponding set of sub-queries for the sources and synthesize a unified global answer for the user.

The original contribution of MOMIS is related to the availability of a set of techniques for the designer to face common problems that arise when integrating pre-existing information sources, containing both semistructured and structured data. MOMIS provides the capability of explicitly introducing many kinds of knowledge for integration, such as integrity constraints, intra- and inter-source intensional and extensional relationships, and designer supplied domain knowledge. A *Common Thesaurus*, which has the role of a shared ontology of the source is built in a semi-automatic way. The *Common Thesaurus* is a set of intra and inter-schema intensional and extensional relationships, describing inter-schema knowledge about classes and attributes of sources schemas; it provides a reference on which to base the identification of classes candidate to integration and subsequent derivation of their global representation.

MOMIS supports information integration in the creation of an integrated view of all sources (Global Virtual View) in a way automated as much as possible and performs revision and validation of the various kinds of knowledge used for the integration. To this end, MOMIS combines reasoning capabilities of Description Logics with affinity-based clustering techniques, by exploiting a common ontology for the sources constructed using lexical knowledge from WordNet and validated integration knowledge.

The Global Virtual View is expressed by using XML standard, to guarantee the interoperability with other open integration system prototype.

3 Integration Process

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

1. *Generation of a Common Thesaurus*.
   The *Common Thesaurus* is a set of terminological intensional and extensional relationships, describing intra and inter-schema knowledge about classes and attributes of sources schemas. we express inter-schema knowledge in form of terminological and extensional relationships (em synonymy, *hypernymy* and *relationship*) between classes and/or attribute names.

2. *Affinity analysis of classes.*
   Relationships in the *Common 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, class structures and relationships in *Common Thesaurus*.

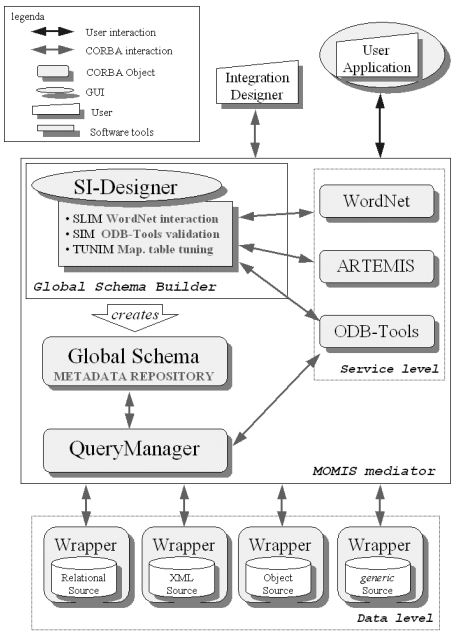

Figure 1: the MOMIS system architecture

3. *Clustering 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 mediated schema*.
   Unification of affinity clusters leads to the construction of the predicted schema. 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 classes derived from clusters, and is the basis for posing queries against the sources.

Global Class and Mapping Tables

Starting from the output of the cluster generation, we define, for each cluster, a *Global Class* that represents the mediated view of all the classes of the cluster. For each global class a set of *global attributes* and, for each of them, the intensional mappings with the *local attributes* (i.e. the attributes of the local classes belonging to the cluster) are given [2].

Shortly, we can say that the global attributes are obtained in two steps: (1) Union of the attributes of all the classes belonging to the cluster; (2) Fusion of the "similar" attributes; in this step redundancies are eliminated in a semi–automatic way taking into account the relationships stored in the *Common Thesaurus*. For each global class a persistent *mapping-table* storing all the intensional

[2]For a detailed description of the mappings selection and of the tool SI-Designer which assist the designer in this integration phase see [2].

```
Vehicle(name, track, length, width, height)
Motor(cod_e, type, compression_ratio, HP,
      KW, lubrification, emission)
Fuel_Consumption(name, cod_e, drive_trains,
      city_km_l, highway_km_l )
Model(name, cod_e, tires, steering, price)
```

Figure 2: Volkswagen database (VW)

```
<!ELEMENT fiat(car*)>
<!ELEMENT car(name,engine,dimensions,tires,
      performance,price)
<!ELEMENT engine(name,cylinders?,layout?,
      capacity_cc?,compression_ratio?,
      power_kw, fuel_system)>
<!ELEMENT dimensions(length,width,heigth,
      luggage_capacity)>
<!ELEMENT performance (urban_consumption,
      combined_consumption,speed)>
<!ELEMENT name (#pcdata)>
...
```

Figure 3: Fiat database (FIAT)

mappings is generated; it is a table whose columns represent the set of the local classes which belong to the cluster and whose rows represent the global attributes. An element $MT[L][ag]$ represents how the global attribute ag is mapped into the local class L. Each element $MT[L][ag]$ of the table can assume one of the following values:

- $MT[L][ag] = al$: the global attribute ag maps into the al local attribute.
- $MT[L][ag] = al_1$ `and` al_2 `and` ... `and` al_n: this is used when the value of the ag attribute is the concatenation of the values assumed by a set of attributes al_i belonging to the same local class L.
- $MT[L][ag]$ = `case of` al $\texttt{const}_1$: $al_1, \ldots$ $\texttt{const}_n$: al_n: this situation occurs when the ag global attribute can assume one value in a set of al_i belonging to the same local class L and the value choice depends on a third attribute, al,from the same class, which act as a selector.
- $MT[L][ag] =$ `const`: in this case a global attribute value does not refer to any local attribute and a constant value is set by the designer (see the Rank attribute).
- $MT[L][ag] =$ `null`: In this case no attribute of the class L corresponds to the global attribute ag.

3.1 Running Example

In order to illustrate how the MOMIS approach works, we will use the following example of integration in the Car manufactoring catalogs, involving two different datasources that collect information about vehicle. The first datasource is the `FIAT` catalog, containing semistructured XML informations about cars of the italian car factory.

The second datasource is the `Volkswagen` database (`VW`), a relational database containing information about this kind of car. Both database schemata are built by analysing the web site of this factory.

3.2 SI-Designer Tool

As described above, the integration process consists of various steps actually implemented in separate module. SI-Designer is a framework that represents a unified solution for the overall integration process.

SI-Designer provides the designer with a graphical interface to reach the Global Virtual View, relating to each integration step a specific interaction with a software module. All the module involved are available as CORBA Object and interact using established `idl` interfaces. In particular the SI-Designer performs this steps:

- **Source acquisition**: in this phase the user can select the sources to be integrated. A wrapper performs the translation from the source description model into ODL_{I^3} description model. This step involves SAM module.
- **Intensional relationships definition**: in this phase, new relationships, *schema derived*, by interacting with SIM module and ODB-Tool system [4], *lexicon derived*, by interacting with the WordNet [12] lexical database, and *designer supplied* are added to the Common Thesaurus (in Figure 4 the relationships involved by our example).

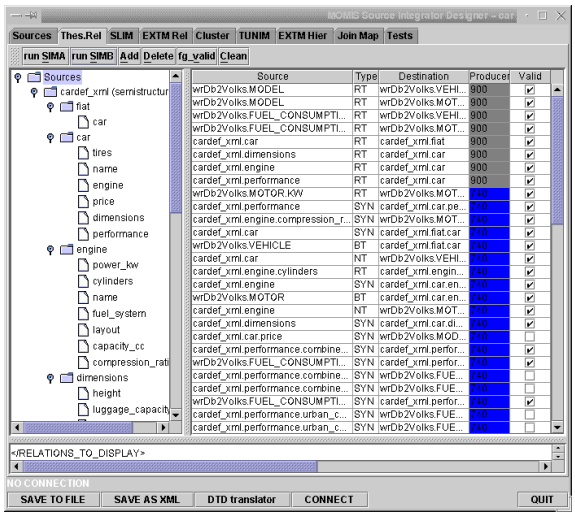

Figure 4: Example: the Common Thesaurus relationships

- **Extensional relationships definition**: Extensional relationships are defined by the interaction with the integration designer. This relationships are exploited to detect extensionally overlapping classes.
- **Clustering**: in this phase, based on the knowledge carried in the Common Thesaurus, by exploiting

ARTEMIS module, global classes are created. In our example (see Figure 5) we obtain mainly a cluster including car data contained in the sources, and a cluster for the motor and engine information.

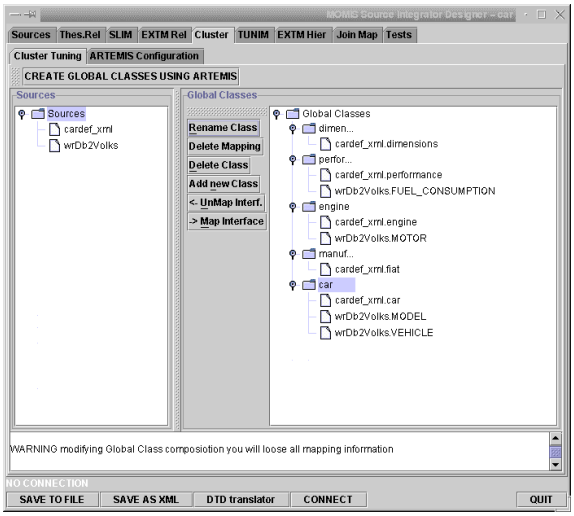

Figure 5: Example: the clustering definition

- **Mapping table tuning**: for each global class generated in the previous phase, the user can modify the Global Virtual View proposed automatically from the system. In Figure 6, Momis system suggests a global class for car rappresentation in which, for example, you may see that global attribute name maps to the related local attributes

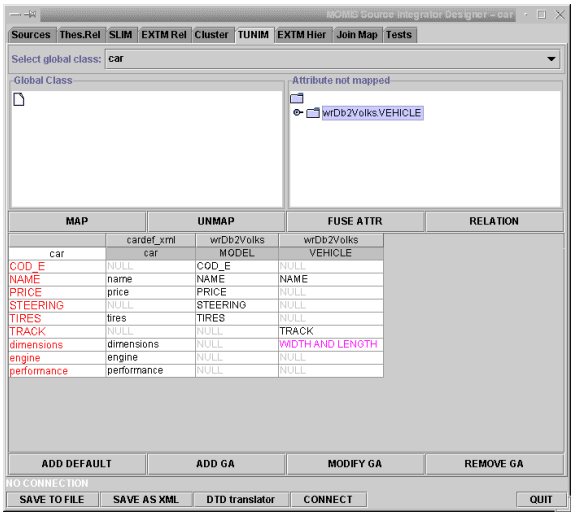

Figure 6: Example: the mapping table tuning

The final step of the integration process provides the export of the Global Virtual View into a XML DTD, by adding the appropriate XML TAGs to represent the mapping table relationships. The use of XML in the definition of the Global Virtual View lets to use Momis infrastructure with other open integration information system by the interchange of XML data files. In addition, the Common Thesaurus is traslated into XML file, so that Momis may provides a shared ontology that can be used by different semantic ontology languages [9, 8].

References

[1] Y. Arens, C.Y. Chee, C. Hsu, and C. A. Knoblock. Retrieving and integrating data from multiple information sources. *International Journal of Intelligent and Cooperative Information Systems*, 2(2):127–158, 1993.

[2] I. Benetti, D.Beneventano, S.Bergamaschi, A. Corni, F. Guerra, and G. Malvezzi. Si-designer: a tool for intelligent integration of information. *International Conference on System Sciences (HICSS2001)*, January 2001. Available at http://www.dbgroup.unimo.it/prototipo/paper/hicss2001.ps.gz.

[3] D. Beneventano, S. Bergamaschi, S. Castano, A. Corni, R. Guidetti, G. Malvezzi, M. Melchiori, and M. Vincini. Information integration: The momis project demonstration. In Amr El Abbadi, Michael L. Brodie, Sharma Chakravarthy, Umeshwar Dayal, Nabil Kamel, Gunter Schlageter, and Kyu-Young Whang, editors, *VLDB 2000, Proceedings of 26th International Conference on Very Large Data Bases, September 10-14, 2000, Cairo, Egypt*, pages 611–614. Morgan Kaufmann, 2000.

[4] D. Beneventano, S. Bergamaschi, C. Sartori, and M. Vincini. ODB-QOPTIMIZER: A tool for semantic query optimization in oodb. In *Int. Conference on Data Engineering - ICDE97*, 1997. http://sparc20.dsi.unimo.it.

[5] S. Bergamaschi, S. Castano, D. Beneventano, and M. Vincini. Semanitc integration of heterogenous information sources. *Journal of Data and Knowledge Engineering*, 36(3):215–249, 2001.

[6] S. Bergamaschi, S. Castano, and M. Vincini. Semantic integration of semistructured and structured data sources. *SIGMOD Records*, 28(1), March 1999.

[7] P. Buneman, L. Raschid, and J. Ullman. Mediator languages - a proposal for a standard, April 1996. Available at ftp://ftp.umiacs.umd.edu/pub/ONRrept/medmodel96.ps.

[8] DAML Joint Committee. Daml Project. Available at http://www.daml.org.

[9] D. Fensel, I. Horrocks, F. van Harmelen, S. Decker, M. Erdmann, and M. Klein. OIL in a nutshell. In *Proceedings of the European Knowledge Acquisition*

Conference (EKAW-2000), Lecture Notes In Artificial Intelligence. Springer-Verlag, 2000. To appear.

[10] Object Management Group. Object management group. http://www.omg.org/.

[11] R. Hull and R. King et al. Arpa i^3 reference architecture, 1995. Available at http://www.isse.gmu.edu/I3_Arch/index.html.

[12] A.G. Miller. Wordnet: A lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

[13] M.T. Roth and 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.