

Information Integration: the MOMIS Project Demonstration

D. Beneventano^{1,*}, S. Bergamaschi^{1,*}, S. Castano², A. Corni^{1,*},

R. Guidetti¹, G. Malvezzi¹, M. Melchiori³, M. Vincini¹

(1) Università di Modena e Reggio Emilia, DSI-Via Campi 213/B, 41100 Modena

(2) Università di Milano, DSI-Via Comelico 39 - 20135 Milano

(3) Università di Brescia DEI-Via Branze 38 - 25123 Brescia

e-mail : {domenico.beneventano,sonia.bergamaschi,corni.alberto,maurizio.vincini}@unimo.it
castano@dsi.unimi.it, melchior@bsing.ing.unibs.it

1 Overview

The goal of this demonstration is to present the main features of a Mediator component, *Global Schema Builder*, of an I3 system, called MOMIS (Mediator environment for Multiple Information Sources) [1]. MOMIS¹² has been conceived to provide an integrated access to heterogeneous information stored in traditional databases (e.g., relational, object-oriented) or file systems, as well as in semistructured sources. The demonstration is based on the integration of two simple sources of different kind, structured and semi-structured, which will be described in Section 2.

Like other integration projects [2, 3], MOMIS follows a “semantic approach” to information integration based on the conceptual schema, or metadata, of the information sources, and on the following functional elements:

1. a common data model, ODM_{I^3} , which is defined according to the ODL_{I^3} language, to describe source

* CSITE-CNR Viale Risorgimento 2, 40136 Bologna

¹Permission to copy without fee all or part of this material is granted provided that the copies are not made or distributed for direct commercial advantage, the VLDB copyright notice and the title of the publication and its date appear, and notice is given that copying is by permission of the Very Large Data Base Endowment. To copy otherwise, or to republish, requires a fee and/or special permission from the Endowment. **Proceedings of the 26th VLDB Conference, Cairo, Egypt, 2000.**

²MOMIS is a joint project among the Università di Modena e Reggio Emilia, the Università di Milano, and the Università di Brescia within the national research project INTERDATA, theme n.3 “Integration of Information over the Web”, coordinated by V. De Antonellis, Università di Brescia.

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 [4]. In addition, ODL_{I^3} introduces new constructors to support the semantic integration process;

2. one or more wrappers, to translate metadata descriptions into the common ODL_{I^3} representation;
3. a mediator which is composed of two modules: the *Global Schema Builder* (GSB) and the *Query Manager* (QM). The GSB 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. In particular, it generates the OQL_{I^3} ³ queries for wrappers, starting from a global OQL_{I^3} query formulated by the user on the global schema. Using Description Logics techniques, the QM component can generate in an automatic way the translation of the generic OQL_{I^3} query into different sub-queries, one for each involved local source.

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

³ OQL_{I^3} is a subset of OQL -ODMG.

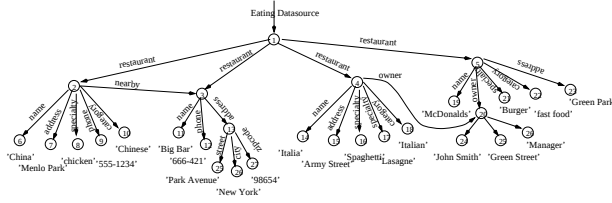


Figure 1: Eating Data Source (ED)

```

Restaurant-pattern =
  (Restaurant,{name, address, phone*,
    specialty, category, nearby*, owner*})
Owner-pattern = (Owner,{name, address, job})
Address-pattern =
  (Address,{street, city, zipcode})

```

Figure 2: Object patterns for the source ED

integration, such as integrity constraints, intra- and inter-source intensional and extensional relationships, and designer supplied domain knowledge. MOMIS supports information integration 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.

2 Demonstration

2.1 Running example

In order to illustrate how the MOMIS approach works, we will use the following example of integration in the Restaurant Guide domain, involving two different datasources that collect information about restau-

```

Steakhouse(s_code, name, street, pers_id,
  special_dish)
Bistro(s_code, type, pers_id)
Person(pers_id, first_name, last_name,
  qualification)
Brasserie(b_code, name, address)

```

Figure 3: Food Guide Database (FD)

rants. The first datasource is the Eating Datasource guidebook (ED), containing semistructured objects about restaurants of the west coast with information about menu, specialties, category, and so on. Fig. 1 illustrates a portion of the ED datasource (with a notation similar to the one of the OEM model [5, 6]).

We use the notion of *object pattern* to represent at intensional level all different objects that describe the same concept in a given semistructured source. Object patterns for all the objects in our semistructured source are shown in Fig. 2 (where the symbol “*” denotes optionality). Three object patterns are defined: Restaurant, containing information about restaurants; Owner, containing information about people involved; Address, containing information about addresses. Each Restaurant has an atomic name, category, and specialty. Furthermore, some Restaurant objects have an atomic address and some other a complex address, a phone, a complex object nearby, specifying the nearest restaurant, and an owner, specifying the name, the address and the job of the restaurant’s owner.

The second datasource is the Food Guide Database (FD), a relational database containing information about USA restaurants. The FD is composed of four relations, namely Steakhouse, Bistro, Person, and Brasserie (see Fig. 3). Information related to restaurant is maintained into the Steakhouse relation. Bistro instance is a subset of Steakhouse instance and contains information about the small informal restaurants that serve wine. Each Steakhouse and Bistro is managed by a Person. Information about places where drinks and snacks are served on, are stored in Brasserie relation.

2.2 Demonstration Architecture

The *Global Schema Builder* (GSB) is the mediator component which processes and integrates ODL_{J3} source descriptions received from wrappers to derive the integrated representation of the datasources, that is, the Global Virtual Schema. It is composed mainly by a GUI (the SI-Designer module), a data repository, and a coordination module (GlobalSchema module) and a set of services (service level) used during the integration (see Figure 4). All such modules are available as CORBA objects and interact using established idl interfaces. Datasources to be

integrated are reachable by *wrapper modules* that are also CORBA objects (with a very simple common interface).

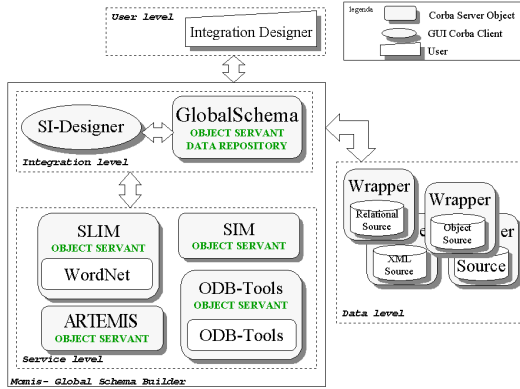


Figure 4: Demonstration architecture

The designer performs the integration process in a semi-automatic way, following the steps suggested by the (**SI-Designer**). Each step is characterized by a graphical form (see Figure 5) and each form “talks directly” with the *GlobalSchema* object (the *idl* interface between *GlobalSchema* and *SI-Designer* is strictly modular) retrieving data and saving new information provided by the designer in the Common Thesaurus, a common ontology among sources.

For the integration phase, GSB uses the following services: **SIM** (*Source Integrator Module*): it extracts intra-schema intensional relationships on the basis of the source structures; **SLIM** (*Schemata Lessical Integrator Module*): it extracts inter-schema intensional relationships between attribute and class names, exploiting the Wordnet lexical system [7]; In this case, synonyms, hypernyms/hyponyms, and related terms can be automatically proposed to the designer, by selecting them according to relationships predefined in the lexical system. **ARTEMIS** tool environment [8], which uses terminological relationships in the Common Thesaurus to assess the level of *affinity* between ODL_{I3} classes by interactively computing a set of affinity coefficients, that take into account both ontological and structural knowledge about ODL_{I3} classes. Furthermore, ARTEMIS classifies ODL_{I3} classes by affinity levels using hierarchical clustering techniques; **ODB-Tools**, a tool based on the OLC Description Logics [9] inference techniques, such as *in-*

coherence detection and *subsumption* computation, which performs ODL_{I3} schema validation and evaluates implicit inter-schema *isa* relationships.

The integration process is subdivided in two phases (1) *Common Thesaurus* generation, (2) Global Virtual Schema generation.

To build the Global Virtual Schema, the following sequence of interactions is performed:

- **SIM**, to extract intra-schema relationships;
- **SLIM**, to extract inter-schema intensional relationships between attribute and class names, exploiting the Wordnet lexical system [7]). At each interaction the extracted relationships are shown to the designer who can confirm them or not and can provide further information.
- **ARTEMIS**, to compute *affinity* coefficients between ODL_{I3} classes, to evaluate their level of matching.

As a result of these interactions, ODB-Tools performs checking and validation to come up with a *Common Thesaurus*. Once the *Common Thesaurus* has been built, the ARTEMIS tool is invoked from within **SI-Designer** to perform clustering and identify clusters of classes with a given *affinity* threshold value. Affinity clusters of ODL_{I3} classes interactively selected by ARTEMIS are passed to ODB-Tools to construct the Global Virtual Schema of the mediator. In particular, an integrated global ODL_{I3} class is interactively defined for each selected cluster. ODB-Tools is exploited for a semi-automatic generation of the global ODL_{I3} classes. The set of global ODL_{I3} classes defined constitutes the Global Virtual Schema of the MOMIS mediator to be used for posing queries against the sources.

The **GlobalSchema** (see Figure 4) is the information repository and acts as coordination object for an integration session. A GlobalSchema object exists for each *integration process*. Such object is characterized by a *status* that spaces between the value *uninitialized* to the value *complete*, corresponding to having a completely modeled Global Virtual Schema. A GlobalSchema object constitutes the input for the *Query Manager* object that will manage queries on the integrated schema.

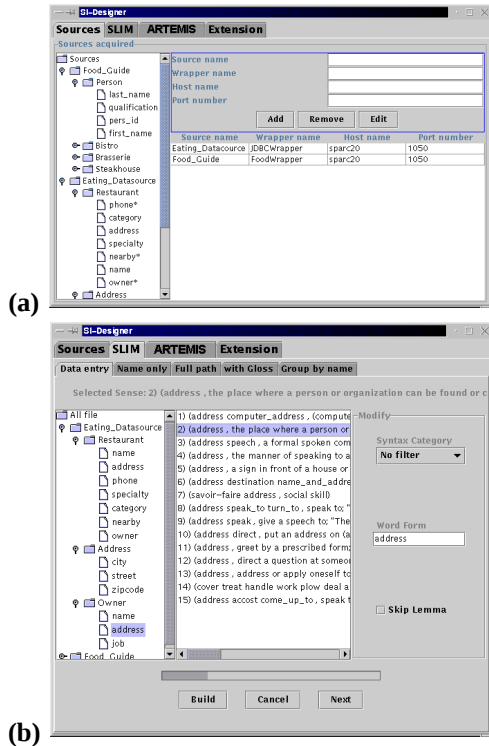


Figure 5: Example: (a) Source binding interface and (b) SLIM interface

References

- [1] S. Bergamaschi, S. Castano, and M. Vincini. Semantic integration of semistructured and structured data sources. *SIGMOD Records*, 28(1), March 1999.
- [2] 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.
- [3] 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.
- [4] P. Buneman, L. Raschid, and J. Ullman. Mediator languages - a proposal for a standard, April

1996. Available at <ftp://ftp.umiacs.umd.edu/pub/ON-Reprt/medmodel96.ps>.

- [5] S. Abiteboul, D. Quass, J. McHugh, J. Widom, and J. Wiener. The lorel query language for semistructured data. *Journal of Digital Libraries*, 1(1), 1996.
- [6] Y.Papakonstantinou, H.Garcia-Molina, and J.Widom. Object exchange across heterogeneous information sources. In *Proc. of ICDE95*, Taipei, Taiwan, 1995.
- [7] A.G. Miller. Wordnet: A lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.
- [8] S. Castano and V. De Antonellis. A schema analysis and reconciliation tool environment for heterogeneous databases. In *IEEE Proc. of IDEAS'99 Int. Database Engineering and Applications Symposium*, Montreal, 1999. ARTEMIS home page: http://bsing.ing.unibs.it/simdeantone/interdata_tema3/Artemis/artemis.html.
- [9] D. Beneventano, S. Bergamaschi, S. Lodi, and C. Sartori. Consistency checking in complex object database schemata with integrity constraints. *IEEE Transactions on Knowledge and Data Engineering*, 10:576–598, July/August 1998.