

IFIP World Computer Congress  
*Topical day:* Semantic Integration of Heterogeneous Data  
Toulouse, Monday 23 August 2004

# Data Integration: general overview and presentation of the MOMIS system

Domenico Beneventano and Sonia Bergamaschi

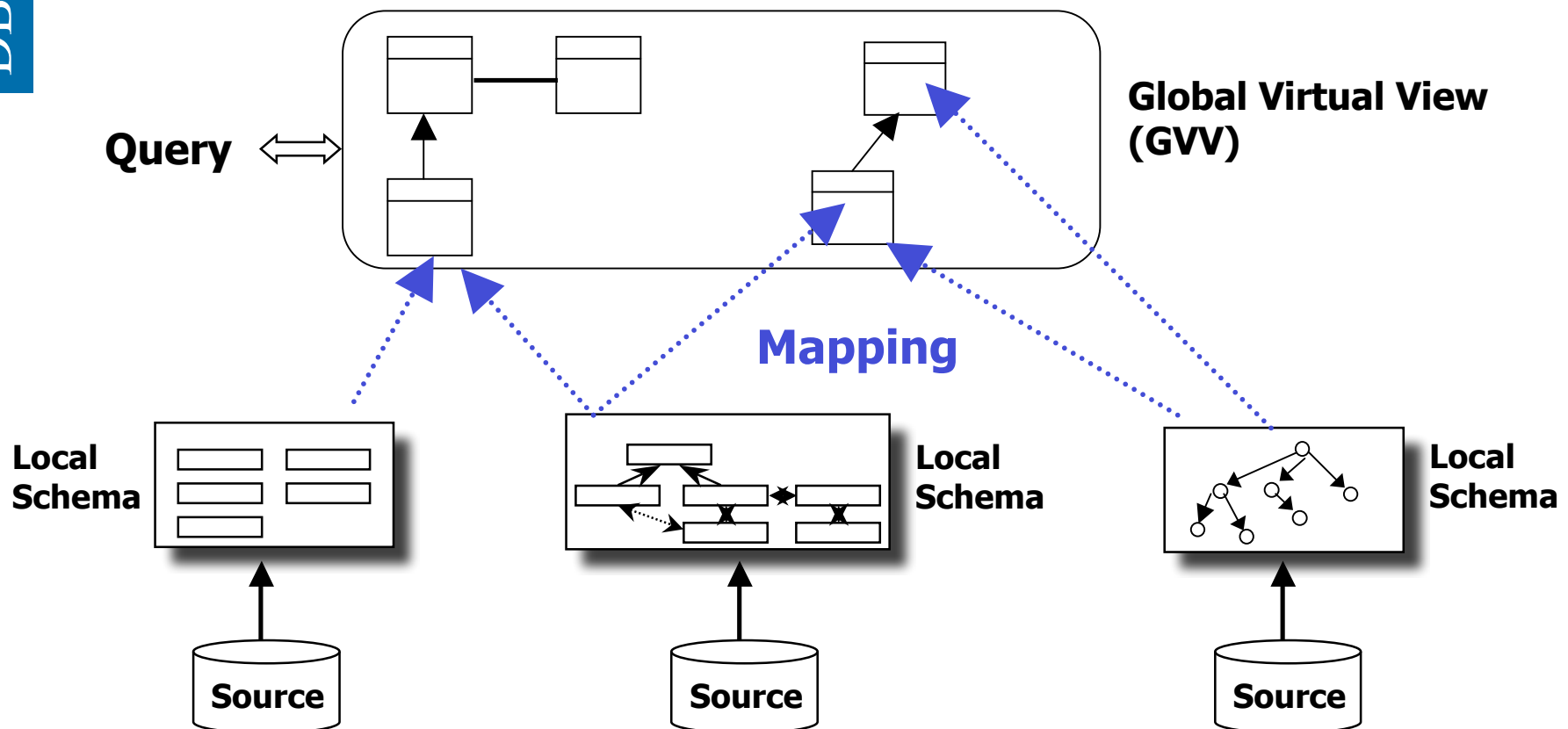
<sup>1</sup>DII- Università di Modena e Reggio Emilia, via Vignolese 905, 41100 Modena

<sup>2</sup>IEEIT-BO V.le Risorgimento 2, 40136 Bologna

- **Semantic Integration of Heterogeneous Data**
- The MOMIS approach to information integration
  - Tool-supported techniques to construct the Global Virtual View
- Global Queries Management
  - Global Queries Management in MOMIS

# Semantic Integration of Heterogeneous Data

- Data integration provides a Global Virtual View (GVV) that
  - is a conceptualization (ontology) describing the involved sources.
  - allows a user to raise a query and to receive a single unified answer



## Main problems in data integration [Lenzerini 2003]

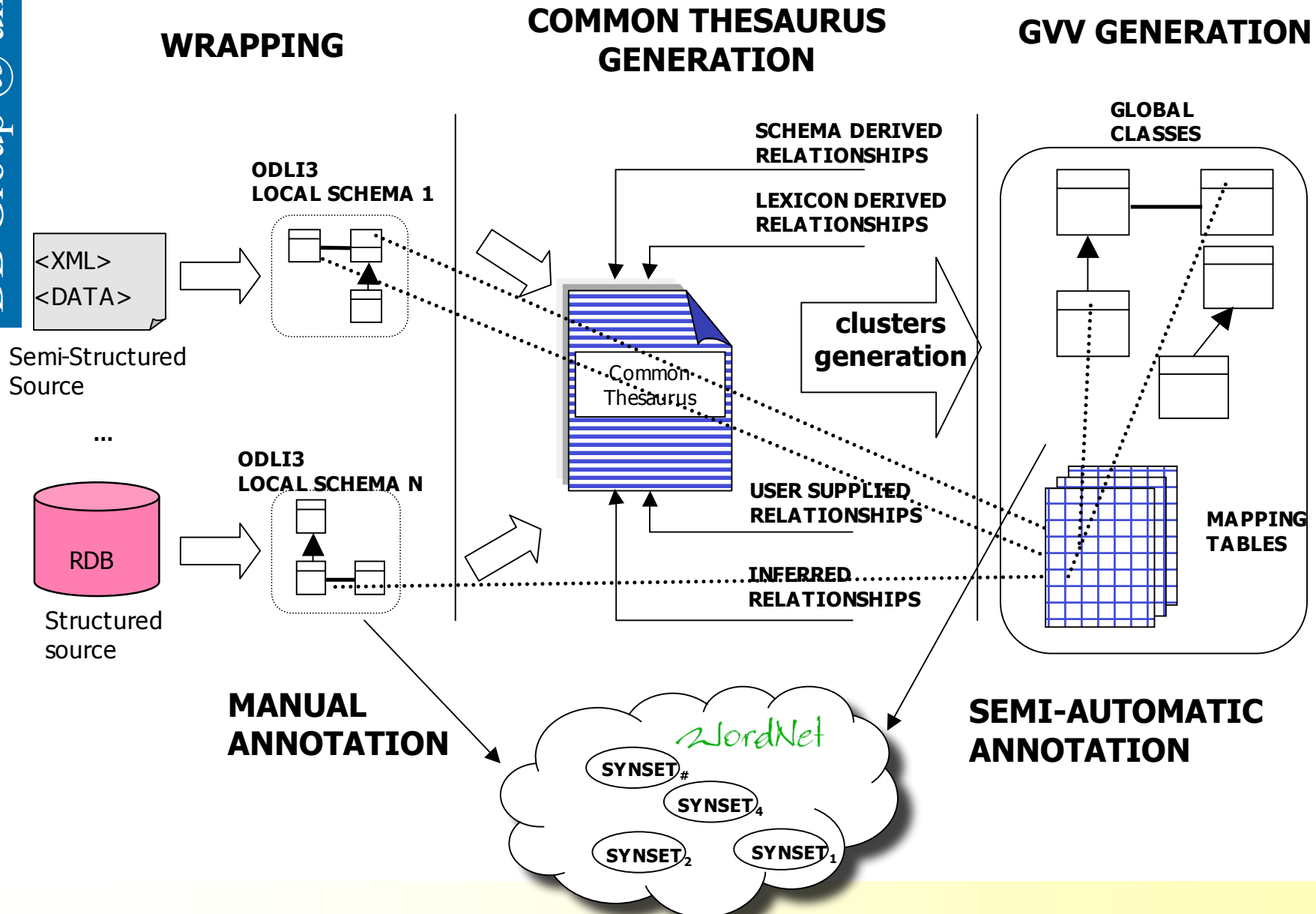
- (Automatic) source wrapping
- How to construct the Global Virtual View
- How to discover interschema properties among the sources and mappings between the sources and the Global Virtual View
- How to model the mappings between the sources and the global virtual view
- How to process updates expressed on the Global Virtual View, and updates expressed on the sources (Schema Evolution)
- How to answer queries expressed on the Global Virtual View (Global Query Management)
- Query optimization
- Data extraction, cleaning and reconciliation (Extensional Integration)

- **MOMIS** (Mediator envirOnment for Multiple Information Sources) is a framework to perform information extraction and integration of heterogeneous, structured and semistructured, data sources
  
- Semantic Integration of Information
  - ❑ A common data model ODLI3 (derived from ODL-ODMG and I3)
  - ❑ The local schema of each source is available (source wrapping)
  
- Tool-supported techniques to construct the Global Virtual View
  - ❑ Local Schema Annotation w.r.t. a common lexical ontology (WordNet)
  - ❑ Semi-automatic generation of relationships between local schemata
  - ❑ Clustering techniques
  - ❑ Semi-automatic generation of mappings between the GVV and local schemata (Mapping Table)
  - ❑ Semi-automatic GVV Annotation w.r.t. a common lexical ontology

- GAV approach: each global class of the GVV is expressed by means of the **full-disjunction** operator [Rajarama, Ullman - 1996]
  
- Query Management over the Global Virtual View
  - Translation (unfolding) of the global query into local queries for the sources
  - Fusion and Reconciliation of the local answers into the global answer
  
- Query Optimization
  - Semantic Query Optimization based on **extensional knowledge**

- Semantic Integration of Heterogeneous Data
- The MOMIS approach to information integration
  - **Tool-supported techniques to construct the GVV**
- Global Queries Management
  - Global Queries Management in MOMIS

# Overview of the GVV-generation process





- UNI (University Source) : XML source represented by a DTD
- CS (Computer Science Source) : relational source

University Source (UNI)	Computer Science Source (CS)
<pre>&lt;!ELEMENT UNI (People*)&gt; &lt;!ELEMENT People (Researcher*, School_Member*)&gt; ... &lt;!ELEMENT Researcher (name, e-mail, Course*, Article*)&gt; &lt;!ELEMENT Teaching ( denomination, specification)&gt; &lt;!ELEMENT Course (name, year, period)&gt; &lt;!ELEMENT Article (title, year, journal, conference)&gt; &lt;!ELEMENT name (#pcdata)&gt; ...</pre>	<pre>Professor (<u>CF</u>, e-mail, first_name, last_name, P_title) FK: P_title references Publication  Student (CF, e-mail)  Class (<u>name</u>, year, description, Prof) FK: Prof references Professor  Publication (<u>title</u>, year, journal, editor)  ...</pre>

- Pieces of the University (UNI) and Computer Science (CS) sources in ODLI3

### UNI Local Schema

```
Interface Researcher
(Source UNI.dtd)
{ attribute string name;
  attribute string e-mail;
  attribute set <Course> courses;
}
Interface Teaching
(Source UNI.dtd)
{ attribute string denomination;
  attribute string description;
}
Interface Course
(Source UNI.dtd)
{ attribute string name;
  attribute integer year;
  attribute string period;
}
...
```

### CS Local Schema

```
Interface Professor
(Source CS.sql)
{ attribute string CF;
  attribute string first_name;
  attribute string last_name;
  attribute string email;
  attribute Publication
                    publication;
Primary Key(CF);
}
Interface Class
(Source CS.sql)
{ attribute string name;
  attribute integer year;
  attribute string description;
  attribute Professor prof;
}
...
```

- Set of intensional and extensional relationships expressing intra-schema and inter-schema knowledge

- **Intensional Relationships**

between class and attribute names (T)

- < T<sub>i</sub> SYN T<sub>j</sub> > *Synonymy*
- < T<sub>i</sub> NT T<sub>j</sub> > *(Narrower Term - NT)*
- < T<sub>i</sub> RT T<sub>j</sub> > *(Related Term - RT)*

- **Extensional Relationships** - between classes (C)

the instances of C1 are ...

- <C1 SYN<sub>Ext</sub> C2> : ... the same instances of C2
- <C1 NT<sub>Ext</sub> C2> : ... a subset of the instances of C2
- <C1 DIS<sub>Ext</sub> C2> : ... disjoint from the instances of C2

- **Common Thesaurus generation:**

- (1) *schema derived relationships*
- (2) *lexicon derived relationships*
- (3) *designer supplied relationships*
- (4) *inferred relationships (exploiting ODB-Tools capabilities)*

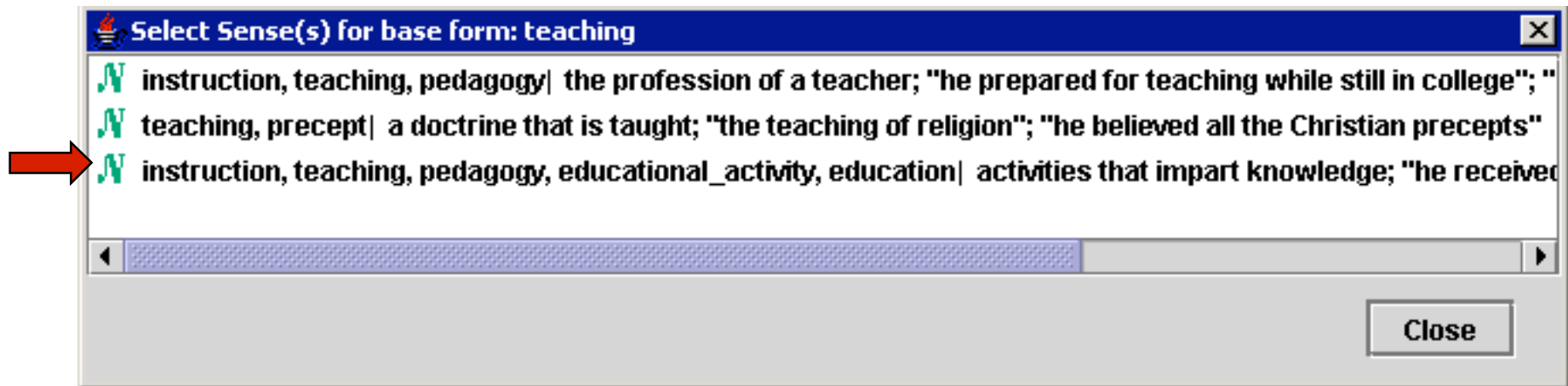
- To assign meanings of class and attribute names with respect to a common lexical ontology (**WordNet**)
- Motivations :
  - to select a well-known meaning for each element of the sources
  - to derive relationships among terms of the sources
- The annotation phase is composed of two steps:
  - 1. Word Form choice.** The WordNet morphologic processor aids the designer by suggesting a word form corresponding to the given term.
  - 2. Meaning choice.** Wordnet proposes the meanings of the term and the designer selects zero, one or more meanings.

# Lexicon-derived Relationships

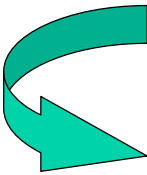
- Extracted from the WordNet lexical ontology
  
- In WordNet:
  - Word forms are organized in synonym set (*synset*)
  - Semantic relationships between *synset* (meanings)
    - Hyponymy (Hypernymy)
    - Meronymy
    - Correlation (between *synsets* having the same Hypernym)
  
- Relationships between class and attribute names are obtained using the WordNet semantic relationships as follows:
  - Synonymy  $\Rightarrow$  SYN
  - Hyponymy  $\Rightarrow$  NT
  - Meronymy and Correlation  $\Rightarrow$  RT

## Local source annotation: Example

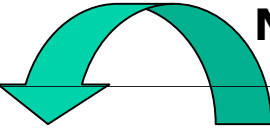
- In the annotation of the class UNI.Teaching the WordNet morphologic processor derives the word form "Teaching" and proposes three meanings:



# Lexicon-derived Relationships : example

**Hyponymy** 

Meaning (synset)	Word Form		
	teaching	course	.. class
education imparted in a series of lessons or class meetings		√	√
activities that impart knowledge	√		
the profession of a teacher			

**NT** 

Common Thesaurus relationships

UNI.COURSE	SYN	CS.CLASS
UNI.COURSE	NT	UNI.TEACHING
CS.CLASS	NT	UNI.TEACHING

# Common Thesaurus Generation: Other rules

- **Schema-derived relationships**

  - RT relationships derived from foreign keys in a relational schema

  - NT relationships from inheritance in a object-oriented schema

  - NT relationships from couples IDs and IDREFs in XML data files

  - ...

- **Inferred relationships**

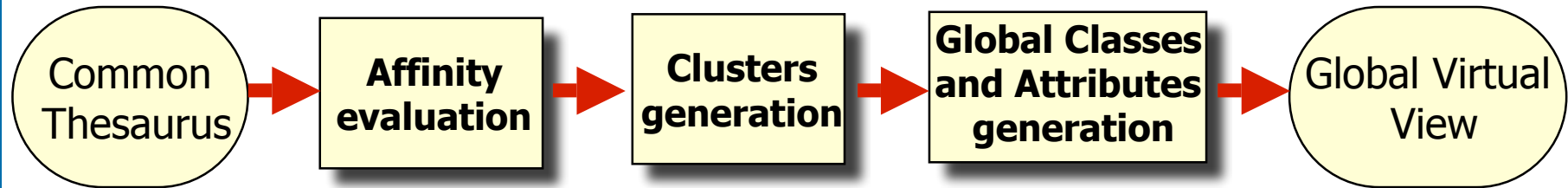
  - Exploiting Description Logics techniques (by using ODB-Tools) a new set of relationships are inferred

- **Designer supplied relationships**

  - The designer can add/delete relationships to the Common Thesaurus



# Global Virtual View Generation



- A global class  $G=(\mathbf{L},\mathbf{GA})$  is generated for each cluster  $C$  :
  - L** are the local classes of the cluster  $C$
  - GA** are the global attributes of  $G$ 
    - Union of the local attributes
    - Fusion of “similar attributes” (by using the Common Thesaurus)

# Mapping Table Generation

How to model the mappings between the local schemata and the GVV?

- Global-As-View (**GAV**) approach:  
the GVV is expressed in terms of the local schemata
- Local-As-View (**LAV**) approach:  
the local schemata are defined in terms of the GVV

- For each global class  $G=(L,GA)$ , a *Mapping Table* (MT) is generated, to represent the mappings between global and local attributes  
MT is a table **GAXL** : An element  $MT[GA][L]$  represents the attributes of the local class  $L$  mapped into the global attribute  $GA$ .
- **Momis** uses a GAV approach where each global class is expressed, on the basis of the Mapping Table, by means of the “**full-disjunction**” [Rajarama, Ullman - 1996] of its local classes.

# GVV and Mapping Table generation : example

- Cluster **G = {UNI.Course, UNI.Teaching, CS.Class}**
- Mapping Table of G

	UNI.Teaching	UNI.Course	CS.Class
<i>Gattribute_1</i>	denomination	name	name
<i>Gattribute_2</i>	description		specification
Year		year	year
Period		period	
Professor			professor

- Since UNI.specification NT CS.description, these local attributes correspond to the same global attribute; the name of this global attribute will be decided in the GVV annotation phase

- Annotating a GVV means to provide each Global Class and each Global Attribute with a name and with a set of meanings with respect to the common lexical ontology.
- We have developed a semi-automatic methodology to generate the annotation of the GVV.

# GVV annotation: example

G = {CS.Class, UNI.Course, UNI.Teaching}

## Annotated Local classes

CS.Class=<class, {class#3}>  
 UNI.Course=<course, {course#1}>  
 UNI.Teaching=<teaching, {teaching#3}>

## Common Thesaurus relationships

UNI.COURSE	SYN	CS.CLASS
UNI.COURSE	NT	UNI.TEACHING
CS.CLASS	NT	UNI.TEACHING

## The annotated Global class

G = <{class, teaching, course}, {class#3, teaching#3, course#1}>

names

broadest name

broadest meaning

meanings

Wordnet  
meanings

class#3 = course#1 = education imparted in a series of  
 lessons or class meetings  
 teaching#3 = activities that impart knowledge

## GVV annotation : example

- A similar approach is used in the annotation of global attributes

Example:

- *Gattribute\_1* ⇒ Name
- *Gattribute\_2* ⇒ Description

- Mapping Table of the global class Teaching

	UNI.Teaching	UNI.Course	CS.Class
Name	denomination	name	name
Description	specification		description
Year		year	year
Period		period	
Professor			professor

- If a class or attribute name has no correspondent in WordNet, the designer may add a new meaning and proper relationships to the existing meanings.
- The designer may add a new meaning (for an existing word-form or for a new one) by:
  - writing the gloss explicitly, or
  - using an existing synset chosen among a list of candidates obtained by an explicit search (using one or more keywords) or by exploiting similarity search techniques.
- The designer may add relationships for the new synset
  - Related synsets are obtained by an explicit search (using one or more keywords) or by exploiting similarity search techniques.

- Semantic Integration of Heterogeneous Data
- The MOMIS approach to information integration
  - Tool-supported techniques to construct the GVV
- **Global Query Management**
  - **Global Query Management in MOMIS**



# Global Query Management

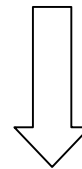
- **The querying problem:** How to answer queries expressed on the GVV (**global queries**)?
- **Query rewriting** : to rewrite a global query as an equivalent set of queries expressed on the local schemata (**local queries**).
  - **GAV** approach: the query is processed by means of **unfolding** (by expanding each atom on the GVV according to its definition in the mapping)
  - **LAV** approach: the query is processed by means of an **inference mechanism** (by re-expressing the atoms on the GVV in terms of the atoms on the local schemata)
- **Fusion and Reconciliation** of the local answers into the global answer
  - **Object Identification**
  - **Inconsistencies** between sources
- **Query Optimization**

# Global Query Management in MOMIS

- **Query rewriting** : MOMIS uses a GAV approach
  - ➔ **Query unfolding** based on the **full-disjunction** operator
  
- **Fusion and Reconciliation** of the local answers into the global answer
  - ➔ **Object Identification : Join conditions** among local classes
  - ➔ **Inconsistencies: Resolution functions** to deal with conflicts
  
- **Query Optimization**
  - ➔ Semantic Query Optimization with **extensional knowledge**

**Local classes  
(relational)**

L1(firstn,lastn,year,e\_mail)  
L2(name,e\_mail,dept\_code,s\_code)



**GVV-Generation**

**Global Class G = { L1, L2}**

**Global Class Schema:**

$S(G) = (\text{Name}, \text{E\_mail}, \text{Year}, \text{Dept}, \text{Section})$

**Global Class Mapping Table:**

	<b>Name</b>	<b>E_mail</b>	<b>Section</b>	<b>Year</b>	<b>Dept</b>
L1	firstn, lastn	e_mail	null	year	null
L2	name	e_mail	s_code	null	dept_code

- Conversion of the local class instances into the GVW instances is performed on the basis of the Mapping Table

	Name	E_mail	Section	Year	Dept
L1	firstn , lastn	e_mail	null	year	null
L2	name	e_mail	s_code	null	dept_code

StringConcatenation(firstn,lastn)

- For each local class L, a **Data Conversion Operator**,  $\delta_L$ , is defined

L1

firstn	lastn	e_mail	year
Rita	Verde	PV@i.it	2
Ada	Rossi	RA@i.it	1

$\delta_{L1}$

Name	E_mail	Year
Rita	PV@i.it	2
Verde Ada Rossi	RA@i.it	1

L2

name	e_mail	dept_c	S_code
Rossi_Ada	RA@i.it	Dept1	413245
Po_Ugo	UP@i.it	Dept1	2314

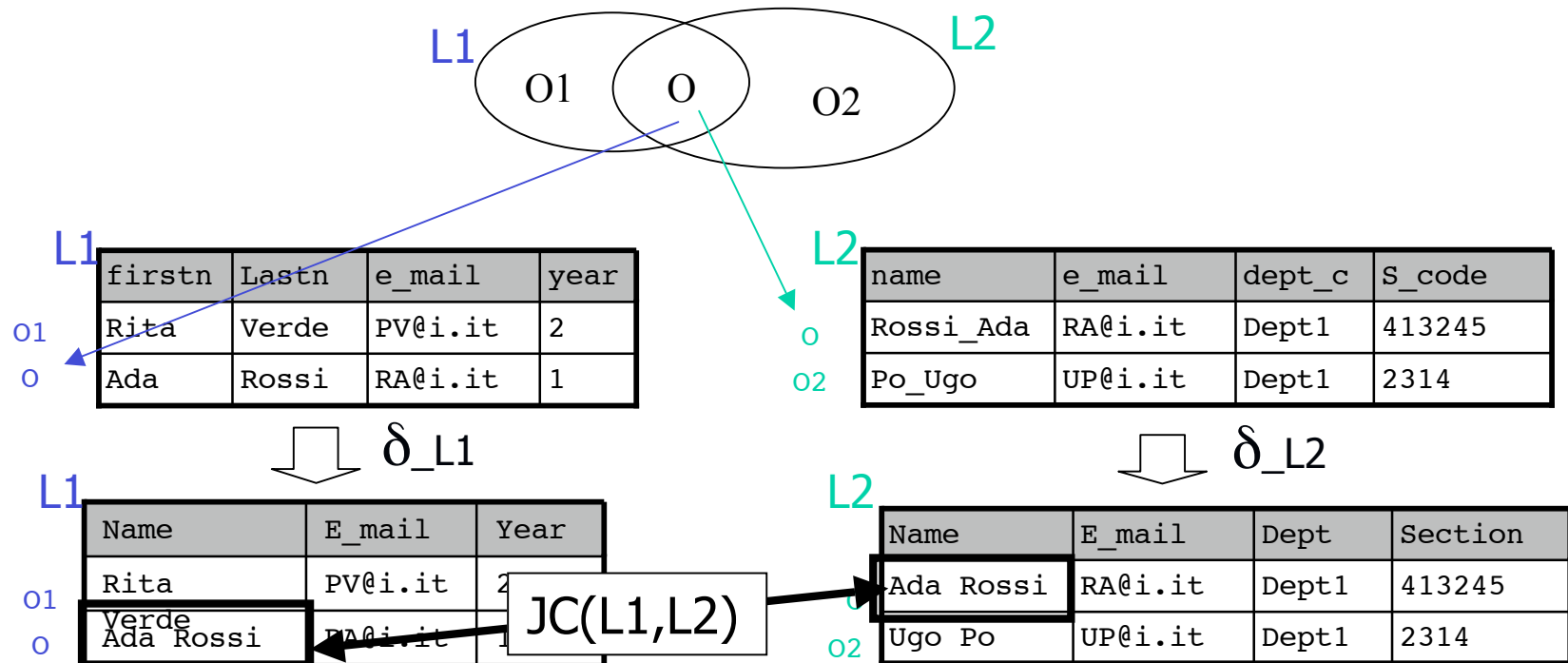
$\delta_{L2}$

Name	E_mail	Dept	Section
Ada Rossi	RA@i.it	Dept1	413245
Ugo Po	UP@i.it	Dept1	2314

# Object Identification

- To identify instantiation of the same object in different sources
  - ➔ Join Conditions among local classes of the same Global Class

Join Condition  $JC(L1,L2) : L1.Name=L2.Name$



## Data Inconsistencies among Sources

- Data stored in the local sources may be **inconsistent** with the **integrity constraints** specified at the global level
- **Example:** data conflict on the `Email` attribute  
⇒ inconsistency with the *Key integrity constraints*

L1

Name	E_mail	Year
Rita Verde	PV@i.it	2
Ada Rossi	ARossi@iol.it	1

L2

Name	E mail	Dept	Section
Ada Rossi	RA@ix.it	Dept1	413245
Ugo Po	UP@i.it	Dept1	2314

## Data Inconsistencies among Sources: different approaches

L1

Name	E_mail	Year
Rita Verde	PV@i.it	2
Ada Rossi	ARossi@iol.it	1

L2

Name	E_mail	Dept	Section
Ada Rossi	RA@ix.it	Dept1	413245
Ugo Po	UP@i.it	Dept1	2314

- Consistent Query Answer: only the consistent data are in the query answer [L. Bertossi, J. Chomicki - 2003]

Name	E_mail
Rita Verde	PV@i.it
Ugo Po	UP@i.it

- Maintaining the conflicts [D. Lembo, M. Lenzerini - 2002]

Name	E_mail
Rita Verde	PV@i.it
Ada Rossi	ARossi@iol.it, RA@ix.it
Ugo Po	UP@i.it

- Resolution Functions to solve the conflicts [F. Naumann - 2000] (MOMIS)

Name	E_mail
Rita Verde	PV@i.it
Ada Rossi	<b>RF</b> (ARossi@iol.it, RA@ix.it)
Ugo Po	UP@i.it

- **Generic resolution function** : Additional input to the resolution function can be values from other domains. For instance, when dealing with different prices, the value of a date attribute might be used to choose the most recent price.
- The **highest informational quality** value on the basis of an information quality model:
  - The quality score can refer to a source in general, or be attribute-specific.
- **Random function**
- Resolution functions for **numerical attributes** : SUM, AVG, ..



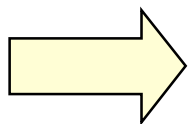
# Homogeneous Attributes

- **No conflicts** : Instances of the same object in different local classes have the same value for common global attributes (Homogeneous Attributes)

	Name	E_mail	Section	Year	Dept
L1	firstn , lastn	e_mail	null	year	null
L2	name	e_mail	s_code	null	dept_code

	Name	E_mail	Year
01	Rita	PV@i.it	2
0	Verde Ada Rossi	RA@i.it	1

	Name	E mail	Dept	Section
0	Ada Rossi	RA@i.it	Dept1	413245
02	Ugo Po	UP@i.it	Dept1	2314



Query Processing with homogeneous attributes

# Full Disjunction

- GAV approach: each global class is expressed by means of the **full-disjunction** of local classes
- Definition of **full-disjunction** [Rajarama, Ullman - PODS 1996]  
 “Computing the natural outerjoin of many relations in a way that preserves all possible connections among facts”
- Given a global class  $G = \{ L1, L2, \dots, Ln \}$ , its instance is the full-disjunction of  $L1, L2, \dots, Ln$  (denoted by  $FD_G(L1, L2, \dots, Ln)$ ) computed on the basis of the Join Conditions

L1

Name	E_mail	Year
Rita	PV@i.it	2
Verde Ada Rossi	RA@i.it	1

L2

Name	E_mail	Dept	Section
Ada Rossi	RA@i.it	Dept1	413245
Ugo Po	UP@i.it	Dept1	2314

$FD_G(L1, L2) : \text{select } S(G) \text{ from } L1 \text{ outer join } L2 \text{ on } JC(L1, L2)$

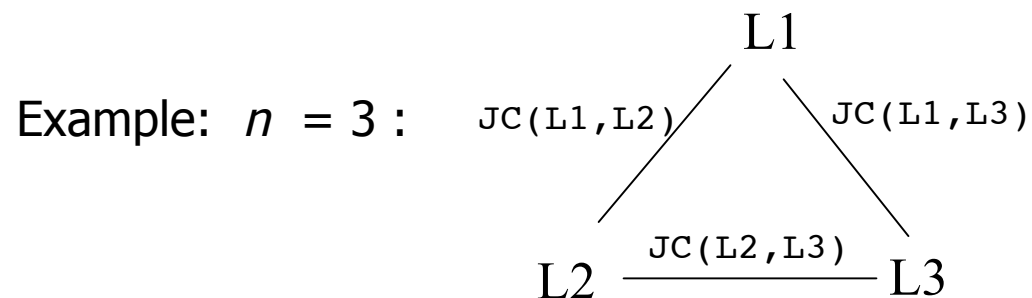
G

Name	E_mail	Year	Dept	Section
Ada Rossi	RA@i.it	1	Dept1	413245
Rita Verde	PV@i.it	2		
Ugo Po	UP@i.it		Dept1	2314

# Full Disjunction Computation

- **Question:** when a full disjunction can be computed by some sequence of natural outerjoins
- **Answer:** there is a natural outerjoin sequence producing the full disjunction if and only if the set of relation schemes forms a connected,  $\gamma$ -acyclic hypergraph [Fagin – 1983]

A Global class with  $n$  local classes,  $n > 2$  :  
 $\gamma$ -cyclic hypergraph



➔ **New Methods**

# Full Disjunction Computation: New Methods

Example:  $n = 3$  :

- *Naive* evaluation

```
select S(G) from
L1 outer join L2 on JC(L1,L2))
outer join L3 on ( JC(L1,L3) OR JC(L2,L3))
```

- outerjoin *pseudo*-sequence

```
select S(G) from
(L1 outer join L2 on JC(L1,L2))
outer join
(L1 outer join L3 on JC(L1,L3))
on JC(L2,L3)
```

# Query rewriting method: UNFOLDING

Global Class  $G = \{ L_1, L_2, \dots, L_n \}$

Global query  $Q$  over  $G$  :  $Q =$  `select <Q_select-list>`  
`from G`  
`where <Q_condition>`

## 1) Local queries

For each local class  $L$ , local query over  $L$  :  $Q_L$

## 2) Full Disjunction of the *converted* local query answers

$Q_{FD} = FD_G (\delta_{L_1}(Q_{L_1}), \dots, \delta_{L_n}(Q_{L_n}))$

## 3) Global query rewriting:

$Q_r =$  `select <Q_select-list>`  
`from  $Q_{FD}$`   
`where <Q_residual-condition>`

## UNFOLDING: local queries computation

Global query **Q** over G :  $Q = \text{select } \langle Q\_select\_list \rangle \text{ from } G$   
 $\text{where } \langle Q\_condition \rangle$

Local query **Q<sub>L</sub>** over L :  $Q\_L = \text{select } \langle Q\_L\_select\_list \rangle \text{ from } L$   
 $\text{where } \langle Q\_L\_condition \rangle$

- 1) **Conjunctive Normal Form of  $\langle Q\_condition \rangle$**
- 2)  **$\langle Q\_L\_condition \rangle$ :**  
constraints of  $\langle Q\_condition \rangle$  which can be solved in L are  
rewritten w.r.t. L (**constraint mapping**)
- 3)  **$\langle Q\_residual\_condition \rangle$ :**  
constraints not included in all local  $\langle Q\_L\_condition \rangle$
- 4)  **$\langle Q\_L\_select\_list \rangle$  :** attributes of the  
 $\langle select\_list \rangle$  of Q + residual constraints + Join Conditions

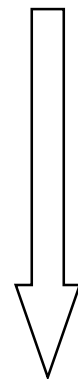
**Hypothesis:**

each global constraint is **fully expressible** at local level

**Example:**

Global constraint:

Name like "Bert\*"



Mapping Table

	Name	...
L1	StringConcatenation(firstn, lastn)	...
L2	name	...

Local constraint for L1:

stringConcatenation(firstn, lastn) like "Bert\*"

Local constraint for L2:

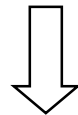
name like "Bert\*"

# Constraint Mapping: other approaches

- In [Chang, García-Molina - 1999], a general framework for Query Mapping across Heterogeneous Information Sources is proposed

## 1. global constraints not expressible at local level

Name like "Bert\*"

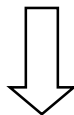


Local constraint for L2

name contains "Bert\*"

## 2. declarative definition of the constraint mapping

Name like "Bert\*"



Local constraint for L1

firstn like "Bert\*" or lastn like "Bert\*"



# Query unfolding example

## Global query

```
Q1:  select E_mail
      from G
      where E_mail like "*.it" and
            (Dept="Dept1" or Year=2)
```

## Local queries

```
Q1_L1: select firstn,lastn,year e_mail
        from L1
        where e_mail like "*.it"
```

```
Q1_L2: select name,dept_c e_mail, from L2
        where e_mail like "*.it"
```

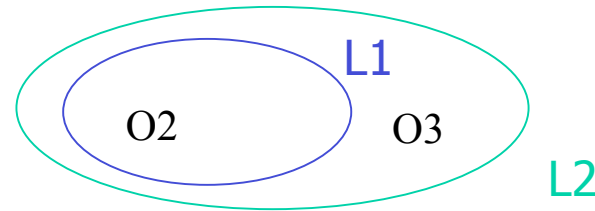
## Global query answer:

```
Q1:  select E_mail
      from  $\delta_{L1}(Q_{L1})$  outer join  $\delta_{L2}(Q_{L2})$  on JM(L1,L2)
      where (Dept= "Dept1" or Year=2)
```

*residual constraint*

# Query unfolding with extensional Knowledge

$L1 \text{ NT}_{\text{EXT}} L2$



Global query

```
Q2:  select e_mail from CG
      where E_mail like '*.it' and Dept = 'Dept1'
```

Local queries

~~Q2\_L1: select firstn, lastn, e\_mail from L1  
where e\_mail like '\*.it'~~

```
Q2_L2: select name, e_mail from L2
        where e_mail like '*.it' and dept_c='Dept1'
```

Global query answer

```
Q2:  select e_mail
      from  $\delta(Q2\_L2)$ 
```

## Conclusion and Future Work

- ✓ We proposed the MOMIS system
  - ✓ to build an integrated GVV of heterogeneous data sources
  - ✓ to answer global queries on the GVV

We are developing MOMIS within the SEWASIE EU-project

- Multilingual issues (EuroWordnet)
  - to consider multilingual information sources
- GVV evolution
  - Update of existing sources
  - Deletion of previously integrated sources
- Global Query Management
  - to consider inconsistent local sources

- SEWASIE (Semantic Webs and AgentS in Integrated Economies) is a research project founded by EU on action line Semantic Web (May 2002/April 2005) <http://www.sewasie.org>
- The consortium details
  - Università degli Studi di Modena e Reggio Emilia (ITALY)
  - CNA SERVIZI Modena s.c.a.r.l. (ITALY)
  - Università degli Studi di Roma “La Sapienza” (ITALY)
  - Rheinisch Westfaelische Technische Hochschule Aachen (GERMANY)
  - Libera Università di Bolzano (ITALY)
  - Thinking Networks AG (GERMANY)
  - IBM Italia SPA (ITALY)
  - Fraunhofer-Gesellschaft zur Förderung der angewandten Forschung eingetragener Verein (GERMANY)



### Contact information:

Prof. Sonia Bergamaschi

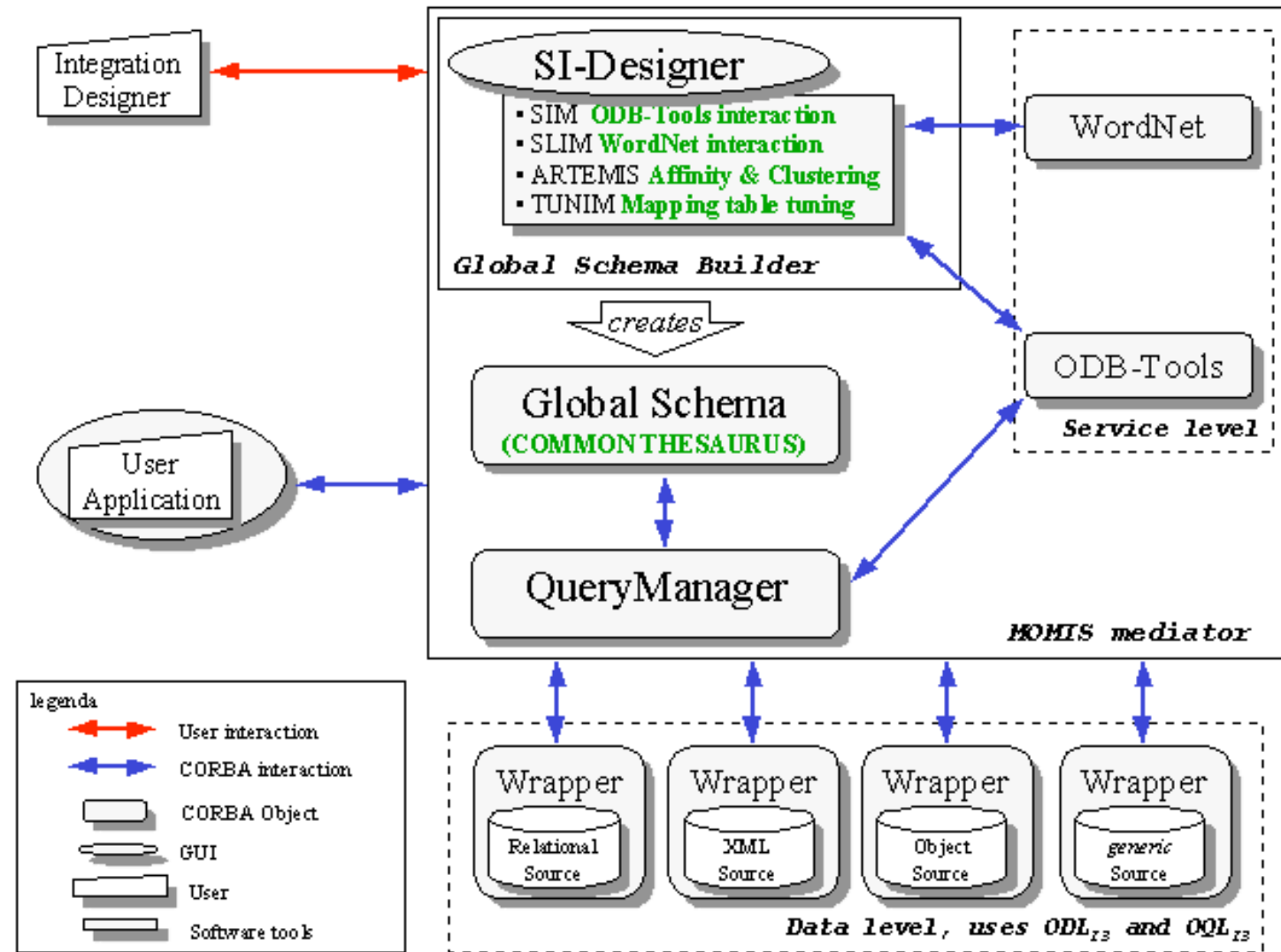
DII – Università di Modena e Reggio Emilia

Tel: +39 059 2056132 Fax: +39 059 2066126

[bergamaschi.sonia@unimo.it](mailto:bergamaschi.sonia@unimo.it)

<http://www.dbgroup.unimo.it/Bergamaschi.html>





Two steps process:

### 1) DTD extraction via a HTML/XML wrapper

HTML/XML wrappers are specialized programs that identify the data of interest in a Web page and map them to the XML format

**Tools:** RoadRunner, Andes, Lixto, ...

### 2) DTD translation into ODLI3

