DB Group @ unimo

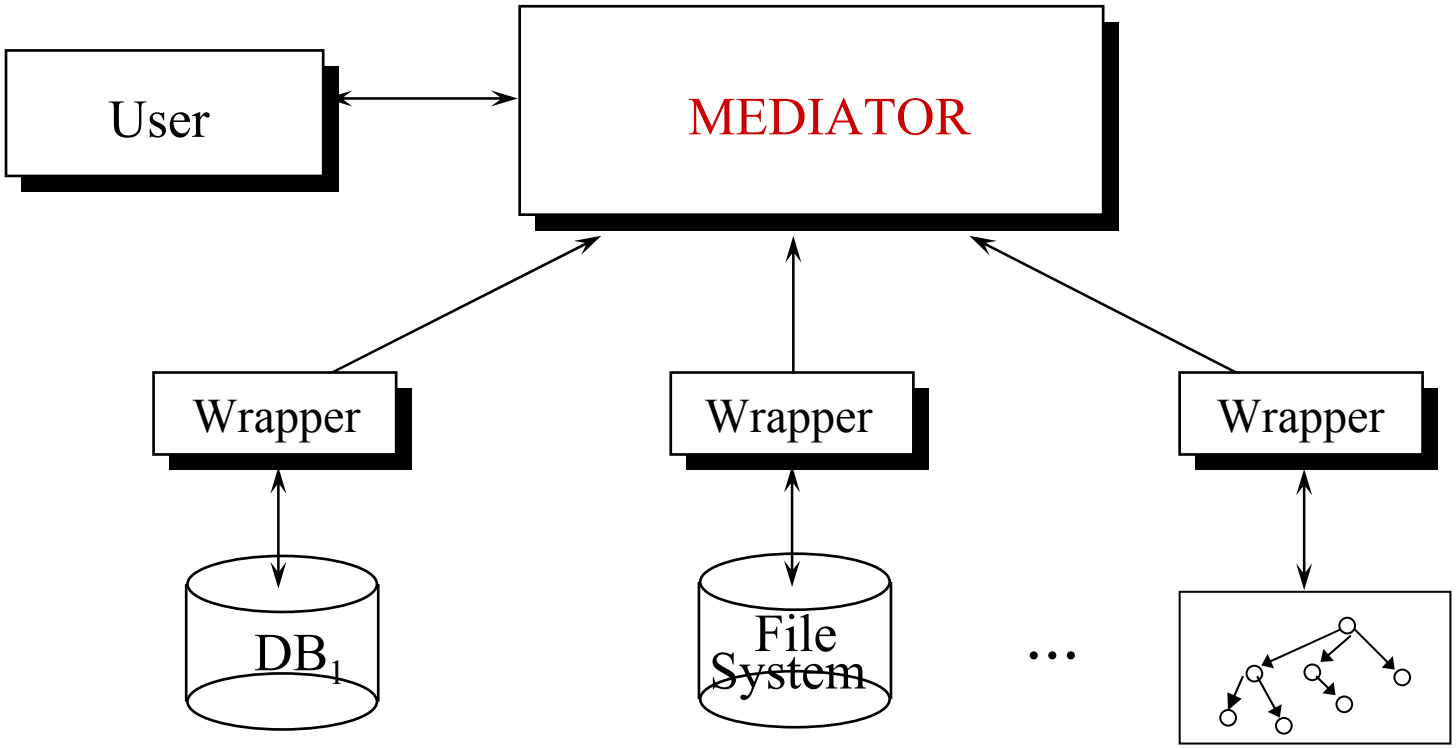
MOMIS:
Mediator envirOnment for Multiple Information Sources

DB-Group http://www.dbgroup.unimo.it

Sonia Bergamaschi, Domenico Beneventano,
Maurizio Vincini, Francesco Guerra, Mirko Orsini,
Laura Po, Antonio Sala

Wrapper/Mediator Architecture

- **Wrapper : extraction of the Local Schemas**
- **Mediator : construction of a Global Virtual View**

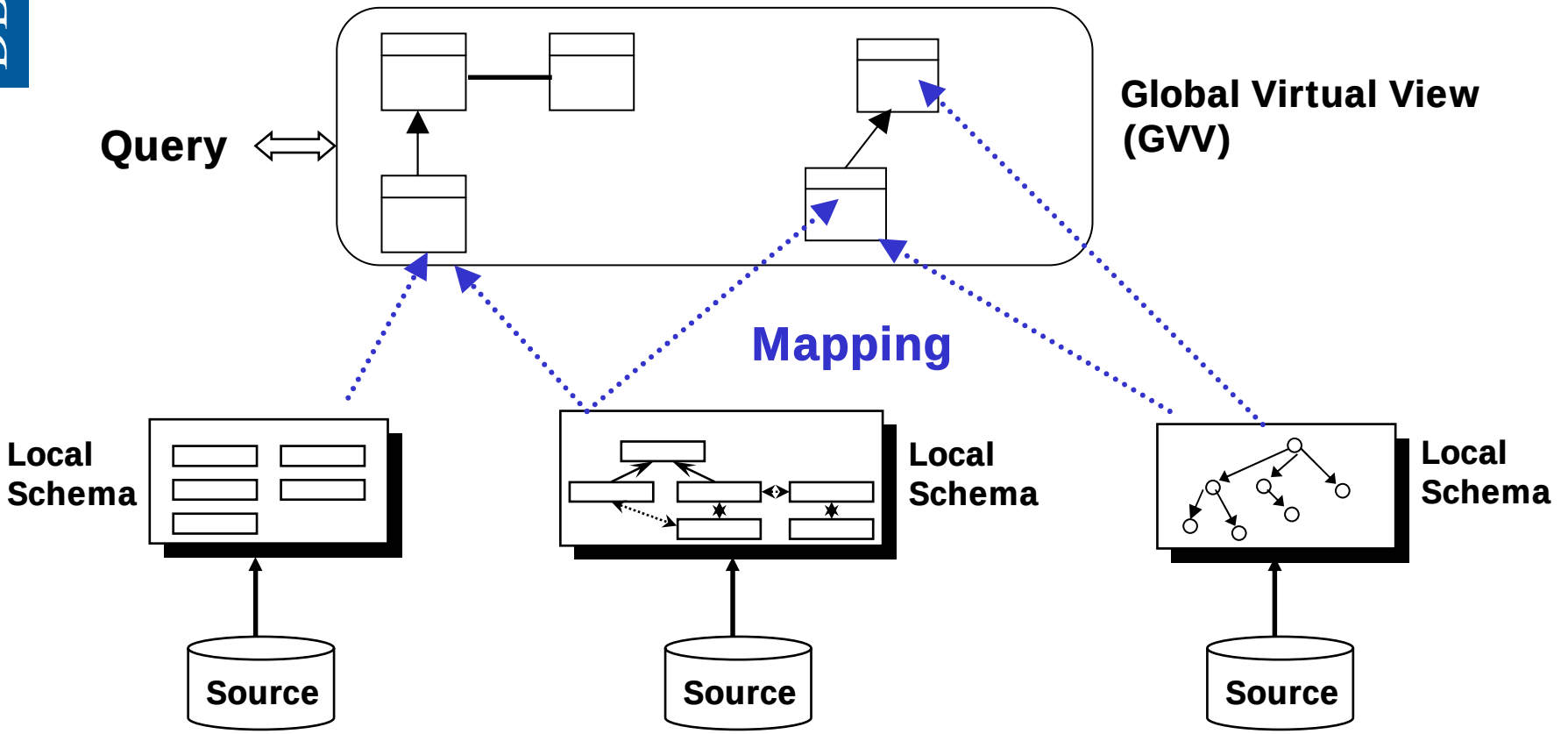


Agenda

- **Semantic Integration of Heterogeneous Data**
- The MOMIS approach to information integration
  - Tool-supported techniques to construct the Global Virtual View
- Global Queries Management
  - Single 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.**
  - development started as a joint collaboration among the University of Modena and Reggio Emilia, the University of Milano and the University of Brescia within INTERDATA (1999-2000); D2I (from Data to Information) (2001-2002) – “Programmi di ricerca scientifica di rilevante interesse nazionale” MIUR.
  - MOMIS development continues at the University of Modena and Reggio Emilia within the IST EU Project SEWASIE(2002-2005)

- Semantic Information Integration
  - A common data model ODLI3 (derived from ODL-ODMG and I3)
  - The local schema of each source is available (source wrapping)

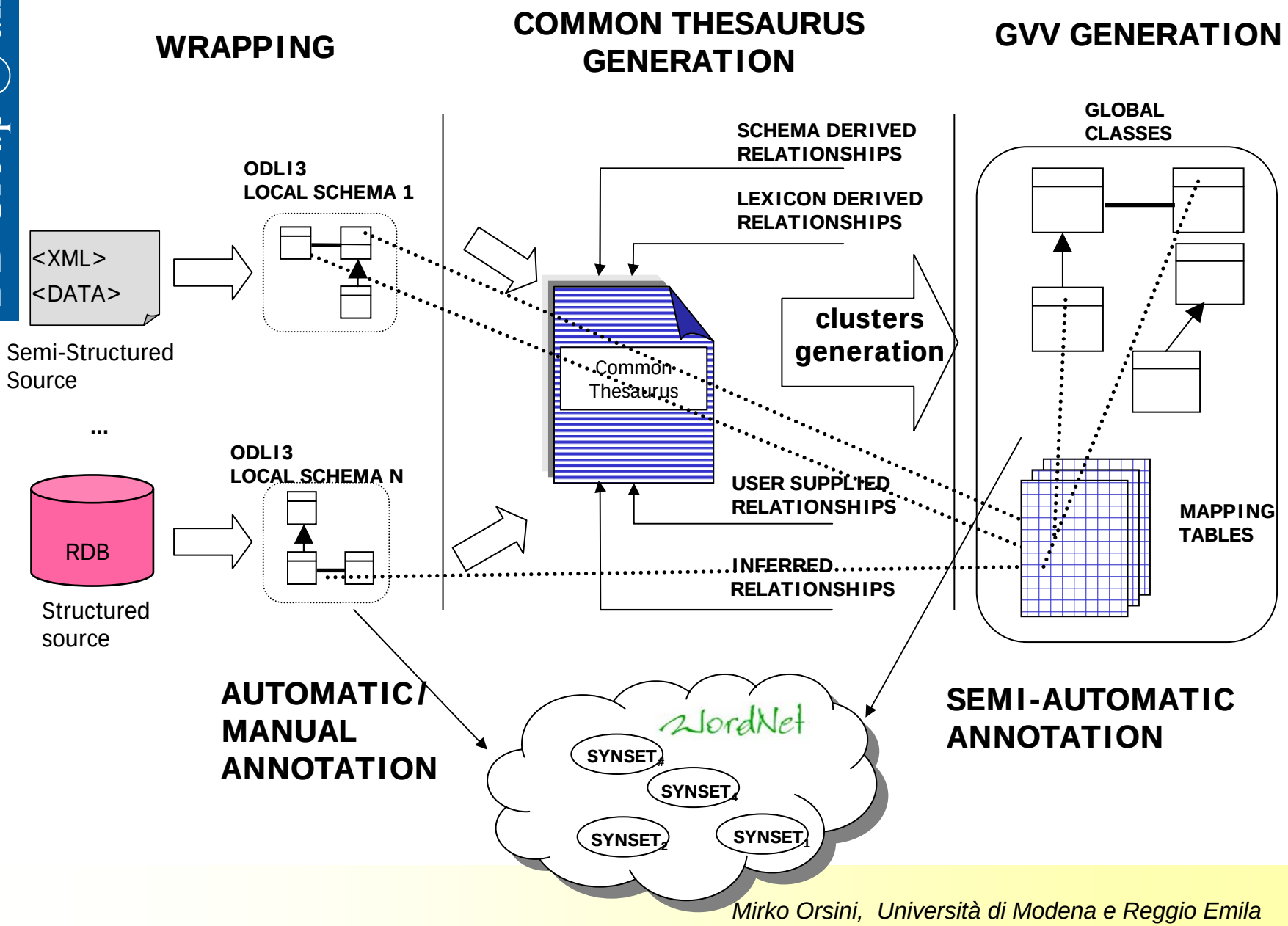
- Tool-supported techniques to construct the Global Virtual View
  - Local Schema Annotation w.r.t. a common lexical ontology (WordNet)
  - Semi-automatic generation of relationships between local schemata
  - Clustering techniques
  - Semi-automatic generation of mappings between the GVV and local schemata (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

- 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



Example (1/2)

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

University Source (UNI)

```
<!ELEMENT UNI(People*)>
<!ELEMENT People(Researcher*,
School_Member*)>
...
<!ELEMENT Researcher(name,
  e-mail,Course*,Article*)>
<!ELEMENT Teaching(
     denomination,
     specification)>
<!ELEMENT Course(name, year,
    period)>
<!ELEMENT Article(title,
  year, journal, conference)>
<!ELEMENT name (#pcdata)> ...
```

Computer Science Source (CS)

```
Professor (CF, e-mail,
   first_name, last_name,
   P_title)
FK: P_title references
            Publication

Student (CF, e-mail)

Class(name,year,
   description, Prof)
FK: Prof references Professor

Publication(title, year,
journal, editor)

…
```

Example (2/2)

- Pieces of the University (UNI) and Computer Science (CS) sources in ODL_{I3}

UNI Local Schema

```
Interface Researcher
(Source UNI.dtd)
{ attribute string name;
  attribute string e-mail;
  attribute set <Course> courses;
 attribute set <Article> articles;
}
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;
}
```

É

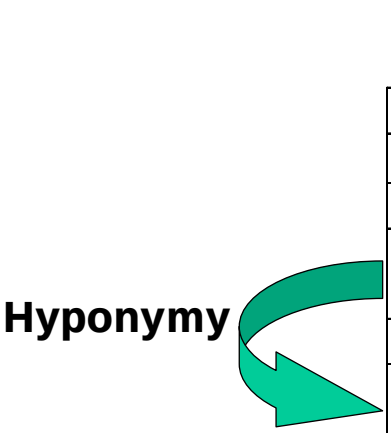
CS Local Schema

```
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;
}
```

É

Annotation and Lexicon-derived Relationships

- **LOCAL SOURCE ANNOTATION** : To assign meanings to class and attribute names w.r.t. a common lexical ontology (**WordNet**)
  - to select a well-known meaning for each element of the sources
  - to derive relationships among elements of the sources



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

Common Thesaurus relationships

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

Extending WordNet

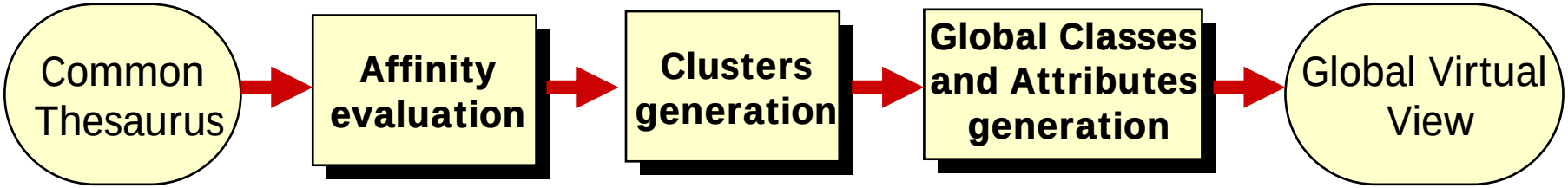
- **WordNet Editor**

- 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.

Global Virtual View Generation



- A global class G=(**L**,**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 **GA**x**L** : 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
Gattribute_1	denomination	name	name
Gattribute_2	description		specification
Year		year	year
Period		period	
Professor			professor

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

GVV Annotation: to provide each Global Class/ Attribute with a name and a set of meanings w.r.t.the common lexical ontology.

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

meanings

broadest meaning

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

- 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 in MOMIS

- **Query rewriting** : MOMIS uses a GAV approach
  - **Query unfolding** based on the **full-disjunction** operator

- **Fusion and Reconcilation** of the local answers into the global answer
  - **Object Identification : Join conditions** among local classes
  - **Inconsistencies: Resolution functions** to deal with conflits

- **Query Optimization**
  - Semantic Query Optimization with **extensional knowledge**