Data Integration: general overview and presentation of the MOMIS system

Domenico Beneventano and Sonia Bergamaschi

1DII- Università di Modena e Reggio Emilia, via Vignolese 905, 41100 Modena
2IEEIIT-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.

Diagram:
- Query
- Global Virtual View (GVV)
- Mapping
- Local Schema
- Source

Domenico Beneventano – Università di Modena e Reggio Emilia
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 thr 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

WRAPPING

COMMON THESAURUS GENERATION

ODL13 LOCAL SCHEMA 1

ODL13 LOCAL SCHEMA N

<XML>
<DATA>

Semi-Structured Source

Structured source

MANUAL ANNOTATION

SEMIAUTOMATIC ANNOTATION

SYNSET_1

SYNSET_2

SYNSET_3

SYNSET_4

USER SUPPLIED RELATIONSHIPS

INFERRED RELATIONSHIPS

SCHEMA DERIVED RELATIONSHIPS

LEXICON DERIVED RELATIONSHIPS

GVV GENERATION

GLOBAL CLASSES

MAPPING TABLES

Domenico Beneventano – Università di Modena e Reggio Emilia
### UNI (University Source) : XML source represented by a DTD

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

**Example (1/2)**
### 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)
    - \(< Ti\ SYN\ Tj >\) Synonymy
    - \(< Ti\ NT\ Tj >\) (Narrower Term - NT)
    - \(< Ti\ RT\ Tj >\) (Related Term - RT)

- **Extensional Relationships** - between classes (C)
  - the instances of C1 are...
    - \(< C1\ SYN_{Ext}\ C2 >\) : ... the same instances of C2
    - \(< C1\ NT_{Ext}\ C2 >\) : ... a subset of the instances of C2
    - \(< C1\ DIS_{Ext}\ C2 >\) : ... 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)
Local sources annotation

- 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 ⇒ SYN
  - Hyponymy ⇒ NT
  - Meronymy and Correlation ⇒ RT
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

<table>
<thead>
<tr>
<th>Word Form</th>
<th>Meaning (synset)</th>
<th>teaching</th>
<th>course</th>
<th>class</th>
</tr>
</thead>
<tbody>
<tr>
<td>education imparted in a series of lessons or class meetings</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>activities that impart knowledge</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>the profession of a teacher</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

## 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
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)
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**
  
  \[ G = \{ \text{UNI.Course}, \text{UNI.Teaching}, \text{CS.Class} \} \]

- **Mapping Table of G**

<table>
<thead>
<tr>
<th>UNI.Teaching</th>
<th>UNI.Course</th>
<th>CS.Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>\textit{Gattribute}_1</td>
<td>denomination</td>
<td>name</td>
</tr>
<tr>
<td>\textit{Gattribute}_2</td>
<td>description</td>
<td></td>
</tr>
<tr>
<td>Year</td>
<td></td>
<td>year</td>
</tr>
<tr>
<td>Period</td>
<td></td>
<td>period</td>
</tr>
<tr>
<td>Professor</td>
<td></td>
<td>professor</td>
</tr>
</tbody>
</table>

- 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.
G = \{\text{CS.Class}, \text{UNI.Course}, \text{UNI.Teaching}\}

Annotated Local classes

- \text{CS.Class} = \langle \text{class}, \{\text{class}\#3}\rangle
- \text{UNI.Course} = \langle \text{course}, \{\text{course}\#1}\rangle
- \text{UNI.Teaching} = \langle \text{teaching}, \{\text{teaching}\#3}\rangle

The annotated Global class

- \text{G} = \langle \{\text{class}, \text{teaching}, \text{course}\}, \{\text{class}\#3, \text{teaching}\#3, \text{course}\#1}\rangle

Common Thesaurus relationships

<table>
<thead>
<tr>
<th>UNI.COURSE</th>
<th>SYN</th>
<th>CS.CLASS</th>
</tr>
</thead>
<tbody>
<tr>
<td>UNI.TEACHING</td>
<td>NT</td>
<td>UNI.TEACHING</td>
</tr>
<tr>
<td>CS.CLASS</td>
<td>NT</td>
<td>UNI.TEACHING</td>
</tr>
</tbody>
</table>

Wordnet meanings

- class\#3 = course\#1 = education imparted in a series of lessons or class meetings
- teaching\#3 = activities that impart knowledge
A similar approach is used in the annotation of global attributes

Example:
- $G\text{attribute}_1 \Rightarrow \text{Name}$
- $G\text{attribute}_2 \Rightarrow \text{Description}$

Mapping Table of the global class Teaching

<table>
<thead>
<tr>
<th>UNI.Teaching</th>
<th>UNI.Course</th>
<th>CS.Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>Name</td>
<td>denomination</td>
<td>name</td>
</tr>
<tr>
<td>Description</td>
<td>specification</td>
<td>description</td>
</tr>
<tr>
<td>Year</td>
<td>year</td>
<td>year</td>
</tr>
<tr>
<td>Period</td>
<td>period</td>
<td></td>
</tr>
<tr>
<td>Professor</td>
<td>professor</td>
<td></td>
</tr>
</tbody>
</table>
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.
Agenda

- 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:

\[ S(G) = (\text{Name}, \text{E}_\text{mail}, \text{Year}, \text{Dept}, \text{Section}) \]

Global Class Mapping Table:

<table>
<thead>
<tr>
<th></th>
<th>Name</th>
<th>E_mail</th>
<th>Section</th>
<th>Year</th>
<th>Dept</th>
</tr>
</thead>
<tbody>
<tr>
<td>L1</td>
<td>firstn, lastn</td>
<td>e_mail</td>
<td>null</td>
<td>year</td>
<td>null</td>
</tr>
<tr>
<td>L2</td>
<td>name</td>
<td>e_mail</td>
<td>s_code</td>
<td>null</td>
<td>dept_code</td>
</tr>
</tbody>
</table>
Conversion of the local class instances into the GVV instances is performed on the basis of the Mapping Table.

For each local class L, a **Data Conversion Operator**, \( \delta_L \), is defined.

- **Data Conversion Operator**
  - For \( L_1 \):
    - Rita Verde PV@i.it 2
    - Ada Rossi RA@i.it 1
  - For \( L_2 \):
    - Rossi Ada RA@i.it Dept1 413245
    - Po Ugo 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

Diagram:

- L1 and L2 are local classes.
- O1 and O2 are objects.
- Join Condition JC(L1,L2) indicates that L1.Name must be equal to L2.Name.
- The diagram illustrates the join process with specific data entries from L1 and L2.

Data Tables:

<table>
<thead>
<tr>
<th>L1 Name</th>
<th>Lastn</th>
<th>E_mail</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rita</td>
<td>Verde</td>
<td><a href="mailto:PV@i.it">PV@i.it</a></td>
<td>2</td>
</tr>
<tr>
<td>Ada</td>
<td>Rossi</td>
<td><a href="mailto:RA@i.it">RA@i.it</a></td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>L2 Name</th>
<th>E_mail</th>
<th>dept_c</th>
<th>S_code</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rossi_Ada</td>
<td><a href="mailto:RA@i.it">RA@i.it</a></td>
<td>Dept1</td>
<td>413245</td>
</tr>
<tr>
<td>Po_Ugo</td>
<td><a href="mailto:UP@i.it">UP@i.it</a></td>
<td>Dept1</td>
<td>2314</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>L1 Name</th>
<th>E_mail</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rita</td>
<td><a href="mailto:PV@i.it">PV@i.it</a></td>
<td></td>
</tr>
<tr>
<td>Ada Rossi</td>
<td><a href="mailto:RA@i.it">RA@i.it</a></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>L2 Name</th>
<th>E_mail</th>
<th>Dept</th>
<th>Section</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ada Rossi</td>
<td><a href="mailto:RA@i.it">RA@i.it</a></td>
<td>Dept1</td>
<td>413245</td>
</tr>
<tr>
<td>Ugo Po</td>
<td><a href="mailto:UP@i.it">UP@i.it</a></td>
<td>Dept1</td>
<td>2314</td>
</tr>
</tbody>
</table>
**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
  \[\Rightarrow\] inconsistency with the *Key integrity constraints*

<table>
<thead>
<tr>
<th>Name</th>
<th>E_email</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rita Verde</td>
<td><a href="mailto:PV@i.it">PV@i.it</a></td>
<td>2</td>
</tr>
<tr>
<td>Ada Rossi</td>
<td><a href="mailto:ARossi@iol.it">ARossi@iol.it</a></td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Name</th>
<th>E_email</th>
<th>Dept</th>
<th>Section</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ada Rossi</td>
<td><a href="mailto:RA@ix.it">RA@ix.it</a></td>
<td>Dept1</td>
<td>413245</td>
</tr>
<tr>
<td>Ugo Po</td>
<td><a href="mailto:UP@i.it">UP@i.it</a></td>
<td>Dept1</td>
<td>2314</td>
</tr>
</tbody>
</table>
Data Inconsistencies among Sources: different approaches

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

<table>
<thead>
<tr>
<th>Name</th>
<th>E_mail</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rita Verde</td>
<td><a href="mailto:PV@i.it">PV@i.it</a></td>
<td>2</td>
</tr>
<tr>
<td>Ada Rossi</td>
<td><a href="mailto:ARossi@iol.it">ARossi@iol.it</a></td>
<td>1</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Name</th>
<th>E_mail</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rita Verde</td>
<td><a href="mailto:PV@i.it">PV@i.it</a></td>
</tr>
<tr>
<td>Ugo Po</td>
<td><a href="mailto:UP@i.it">UP@i.it</a></td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Name</th>
<th>E_mail</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rita Verde</td>
<td><a href="mailto:PV@i.it">PV@i.it</a></td>
</tr>
<tr>
<td>Ada Rossi</td>
<td><a href="mailto:ARossi@iol.it">ARossi@iol.it</a>, <a href="mailto:RA@ix.it">RA@ix.it</a></td>
</tr>
<tr>
<td>Ugo Po</td>
<td>RF(<a href="mailto:ARossi@iol.it">ARossi@iol.it</a>, <a href="mailto:RA@ix.it">RA@ix.it</a>)</td>
</tr>
</tbody>
</table>
Resolution Functions

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

<table>
<thead>
<tr>
<th>Name</th>
<th>E_mail</th>
<th>Section</th>
<th>Year</th>
<th>Dept</th>
</tr>
</thead>
<tbody>
<tr>
<td>L1</td>
<td>firstn, lastn</td>
<td>e_mail</td>
<td>null</td>
<td>year</td>
</tr>
<tr>
<td>L2</td>
<td>name</td>
<td>e_mail</td>
<td>s_code</td>
<td>null</td>
</tr>
</tbody>
</table>

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 \( G = \{ L_1, L_2, \ldots, L_n \} \), its instance is the full-disjunction of \( L_1, L_2, \ldots, L_n \) (denoted by \( \text{FD}_G(L_1, L_2, \ldots, L_n) \)) computed on the basis of the Join Conditions

\[
\text{FD}_G(L_1, L_2) : \text{select } S(G) \text{ from } L_1 \text{ outer join } L_2 \text{ on } JC(L_1, L_2)
\]
• **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

Example: \(n = 3\):

\[
\begin{array}{c}
\text{L1} \\
\downarrow \text{JC(L1,L2)} \\
\text{L2} \\
\downarrow \text{JC(L2,L3)} \\
\text{L3}
\end{array}
\]

New Methods
Example: $n = 3$:

- \textit{Naive evaluation}

\begin{verbatim}
select S(G) from 
L1 outer join L2 on JC(L1,L2)) 
outer join L3 on ( JC(L1,L3) OR JC(L2,L3))
\end{verbatim}

- \textit{outerjoin pseudo-sequence}

\begin{verbatim}
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)
\end{verbatim}
Global Class $G = \{ L_1, L_2, ..., L_n \}$

Global query $Q$ over $G$: 

\[
Q = \text{select} \ <Q\text{-select-list}> \\
\text{from} \ G \\
\text{where} \ <Q\text{-condition}>
\]

1) **Local queries**

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

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

\[
Q_{FD} = FD_G (\delta_{L_1}(Q_{L_1}), ..., \delta_{L_n}(Q_{L_n}))
\]

3) **Global query rewriting**:

\[
Q_r = \text{select} \ <Q\text{-select-list}> \\
\text{from} \ Q_{FD} \\
\text{where} \ <Q\text{-residual-condition}>
\]
UNFOLDING: local queries computation

Global query $Q$ over $G$:
$$Q = \text{select } <Q\_select\_list> \text{ from } G \text{ where } <Q\_condition>$$

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

1) **Conjunctive Normal Form of** $<Q\_condition>$

2) $<Q\_L\_condition>$:
   constraints of $<Q\_condition>$ which can be solved in $L$ are rewritten w.r.t. $L$ *(constraint mapping)*

3) $<Q\_residual\_condition>$:
   constraints not included in all local $<Q\_L\_condition>$

4) $<Q\_L\_select\_list>$: attributes of the $<select\_list>$ of $Q$ + residual constraints + Join Conditions
Hypothesis: each global constraint is **fully expressible** at local level

Example:

Global constraint:

Name like "Bert*"

Local constraint for L1:

stringConcatenation(firstn, lastn) like "Bert*"

Local constraint for L2:

name like "Bert*"

Mapping Table

<table>
<thead>
<tr>
<th>Name</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>L1</td>
<td>...</td>
</tr>
<tr>
<td>L2</td>
<td>...</td>
</tr>
</tbody>
</table>

Domenico Beneventano – Università di Modena e Reggio Emilia
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*"
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 δ_L1(Q_L1) outer join δ_L2(Q_L2) on JM(L1, L2)
    where (Dept= "Dept1" or Year=2)

residual constraint
Query unfolding with extensional Knowledge

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 δ(Q2_L2)
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

• 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 Westfälische 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
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**