The MOMIS approach for Information Integration

Sonia Bergamaschi
www.dbgroup.unimo.it
Dipartimento di Ingegneria “Enzo Ferrari”
Università di Modena e Reggio Emilia, via Vignolese 905, 41100 Modena

www.datariver.it
Spin-off presso Università di Modena e Reggio Emilia
• Modern enterprises are often organized as “virtual networks”, where enterprises operate through inter-enterprise cooperative processes.

• To manage inter-enterprise processes and data exchange a key issue is to mediate among the heterogeneity of different information systems: **Data Integration is a technological solution to build a shared and integrated knowledge base.**

• Data Integration has to deal with the problems arising from the heterogeneity of data sources:

  • **Structural Heterogeneity:**
    – Different data models
    – Same model but different conceptualization chosen

  • **Semantic Heterogeneity:**
    different meaning and interpretation
    – Two schemata might use the same term to denote distinct concepts (homonyms), or different terms to denote the same concept (synonyms)
A **Mediated Schema** provides an integrated and virtual view of the data stored in several sources. No centralized copy of data is stored, a user query on the mediated schema is transformed into queries over the original sources.
Data from several sources are extracted, transformed and loaded into a Data Warehouse, where users can run queries.
Data integration provides a Global *virtual* Schema (GS) that
- is a conceptualization (ontology) describing a set of distributed & heterogeneous data sources
- allows a user to pose a query and receive a unique answer (transparently from the involved sources)
Virtual Data Integration: an example

Search mail and phone of professors whose research activities are in the “Database area”

Select mail, phone From Faculty_member Where research_topic = “Database”

Select e-mail From Professor Where area = “Database”

Global Schema

Professor

name phone mail area

Faculty Member

name phone mail research_topic

Local Schemata

Source 1

Professor

fname lname e-mail area

Source 2

Faculty Member

name phone mail research_topic
**MOMIS** (Mediator envirOnment for Multiple Information Sources) is a framework to perform information extraction and integration of heterogeneous, structured and semistructured, data sources developed by the DBGroup at the University of Modena and Reggio Emilia (www.dbgroup.unimo.it/Momis)

**Semantic Integration of Information**
- A common data model ODL$_{I^3}$ (derived from ODL-ODMG and I$^3$)
- The local schema of each source is available (source wrapping)

**Tool-supported techniques to construct the Global Schema**
- Local Schema Annotation w.r.t. a common lexical ontology (WordNet)
- Semi-automatic generation of mappings among local schemata
- Clustering techniques
- Semi-automatic generation of GAV (Global as View) mappings between the GS and local schemata (Mapping Table)

**Global Query Management**
The **DataBase Group** ([www.dbgroup.unimo.it](http://www.dbgroup.unimo.it)) is the research database group at the Department of Computer Engineering of the University of Modena and Reggio Emilia, it is led by Professor Sonia Bergamaschi and is composed of the following researchers:

- Sonia Bergamaschi (full professor)
- Domenico Beneventano (professor)
- Maurizio Vincini (professor)
- Francesco Guerra (senior researcher)
- Mirko Orsini (Phd – CEO of DATARIVER)
- Laura Po (Phd - researcher)
- Antonio Sala, Serena Sorrentino (Phd - research collaborator)
- Fabio Benedetti, Giovanni Simonini (Phd student)
- Silvia Rota, Dannaoui Abdul Rahman (Phd students)
- Alberto Corni (Phd - research collaborator)
DB Group research activity

- **Big Data Management & Analysis**
  - NOSQL DBMS
  - Business Intelligence

- **Intelligent Information Integration**
  - to combine data residing at different autonomous sources, and providing the user with a unified view of these data

- **Semantic Search Engines**
  - to augment and improve traditional Web Search Engines by using not just words, but concepts and semantic relationships
  - Keyword Search on Relational Databases
  - Recommendation Systems
National and International Research Projects

- Project Partecipation: “D2I (From Data to Information)” supported by MIUR: “Programma di ricerca scientifica di rilevante interesse nazionale (2000-2001)”;

- Project Partecipation: “Agenti software e commercio elettronico: profili giuridici, tecnologici e psico-sociali”, supported by MIUR “Programma di ricerca scientifica di rilevante interesse nazionale” (2001-2002)

- Project Partecipation: “Tecnologie per arricchire e fornire accesso a contenuti” supported by MIUR - Fondo Speciale Innovazione 2000 (2001-2002)

- Project Participation: “CROSS “ supported by Regione Emilia-Romagna Iniziativa 1.1 PRRIITT(September 2005-2007)

National and International Research Projects

- Project Participation: “STIL“ supported by Regione Emilia-Romagna Iniziativa 1.1 del Piano Telematico Regionale (September 2005-2007)
- Project Coordination: “SEWASIE (SEmantic Web and AgentS in Integrated Economies)” supported by IST-UE RDT(2002-2005)
- Project Coordination: “WISDOM (Web Intelligent Search based on DOMain ontologies)” supported by MIUR “Programma di ricerca scientifica di rilevante interesse nazionale” (2005-2007)
- Project Participation: “STASIS (SofTware for Ambient Semantic Interoperable Services)” (2006-2008) supported by IST-EU RDT - started on september 2006
Project Participation: BIOGEST-SITEIA lab (2011-2013)

Project Coordination: “KEYMANTIC” founded by the “Fondazione Cassa di Risparmio di Modena (http://www.fondazione-crm0.it/), whose aim is the development of a keyword-based query engine supporting users in querying data sources with complex and large schemas (2009-2011)

Project Participation: “FACILITATE (SofTware for Ambient Semantic Interoperable Services)” (2010-2012) supported by IST-EU RDT
MOMIS has been already tested in the above mentioned research projects for the development of Vertical Web Portals and the integration of heterogeneous data sources in many domains:

- Tourism (vertical web portal - WISDOM)
- Textile (search engine - SEWASIE)
- Mechanical (search engine - SEWASIE)
- Logistics (logistic domain ontology - STIL)
- Agro-Food (data integration for cereals breeding - CEREALAB)
- Commercial (business intelligence - CROSS)

MOMIS provides methods and tools for:

- sharing legacy systems in an integrated information system
- safeguarding the autonomy of systems and organizations
- support the enterprise interoperability
In the **WISDOM** project (**Web Intelligent Search based on DOMAIN ontologies**) (www.dbgroup.unimo.it/wisdom) the **MOMIS** system was exploited for the integration of several tourism web sites and the development of a Tourism Vertical Web Portal.
In the **SEWASIE** project (**SEmantic Web and AgentS in Integrated Economies**) (www.sewasie.org) the **MOMIS** system was exploited for the integration of heterogeneous company data sources and the development of a Semantic Search Engine.
In the **STIL** project (www.stil-project.org) the **MOMIS** system was exploited for the integration of logistic company data sources and the development of a logistics domain ontology. The “Virtual Logistic Hub” based on the logistics domain ontology provides interoperability between logistic enterprises.
In the **CEREALAB** project (www.cerealab.unimore.it) the **MOMIS** system was exploited for the integration of molecular and phenotypic data sources and the development of an integrated information system for cereals breeders.

Now adopted in the Biogest-Siteia project.
In the **CROSS** project (www.cross-lab.it) the **MOMIS** system was exploited for the integration of heterogeneous sales order data sources and the population of a Data Warehouse in a business intelligence environment.
MOMIS Architecture

Integration Designer

User Application

SI-Designer
- SLIM: WordNet interaction
- SIM: ODB-Tools validation
- TUNIM: Mapping table tuning

Global Schema Builder

WordNet
ARTEMIS
ODB-Tools

Service level

Global Schema Metadata repository

Query Manager

User interaction
Corba interaction

MOMIS Mediator

Wrapper
Wrapper
...

DB Group @ unimo
1. Global Schema and Mapping generation

⇒ the Global Schema is generated and its mappings with the data source are defined

2. Querying the Integrated Data

⇒ the integrated data can be queried by users and software applications
Overview of the GS generation process

WRAPPING

<XML>
<Data>
Semi-Structured Source

... 

RDB
Structured source

COMMON THESAURUS GENERATION

ODLI3 LOCAL SCHEMA 1

ODLI3 LOCAL SCHEMA N

AUTOMATIC/MANUAL ANNOTATION

GLOBAL CLASSES

GS GENERATION

SCHEMA DERIVED RELATIONSHIPS

LEXICON DERIVED RELATIONSHIPS

USER SUPPLIED RELATIONSHIPS

INFERRED RELATIONSHIPS

clusters generation

SEMIAUTOMATIC ANNOTATION & OWL exportation

LEXICON DERIVED RELATIONSHIPS

SCHEMA DERIVED RELATIONSHIPS

USER SUPPLIED RELATIONSHIPS

INFERRED RELATIONSHIPS

clusters generation

SEMIAUTOMATIC ANNOTATION & OWL exportation
Common Thesaurus

- 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)
Lexicon-derived Relationships

- Extracted from the WordNet thesaurus (lexical ontology)

- In WordNet:
  - Word forms are organized in synonym sets (*synsets*)
  - Semantic relationships between *synsets* (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
Hyponymy (is a kind of)

<table>
<thead>
<tr>
<th>Meaning (synset)</th>
<th>Book</th>
<th>Volume</th>
<th>Publication</th>
</tr>
</thead>
<tbody>
<tr>
<td>a written work or composition that has been published (printed on pages bound together)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>physical objects consisting of a number of pages bound together; &quot;he used a large book as a doorstop&quot;</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>the amount of 3-dimensional space occupied by an object</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>a copy of a printed work offered for distribution</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Lexicon derived relationships

<table>
<thead>
<tr>
<th>Book</th>
<th>SYN</th>
<th>Volume</th>
<th>Publication</th>
</tr>
</thead>
<tbody>
<tr>
<td>Book</td>
<td>NT</td>
<td></td>
<td>Publication</td>
</tr>
</tbody>
</table>
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.
**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 and Mapping Table Generation

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

- **MT generation**:  
  For each global class $C=(L, GA)$, 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 $GA$. 
GS and MT generation: example

Cluster

\[ \text{Company}=\{\text{prontocomune.Azienda, fibre2fashion.Company, usawear.Company}\} \]

Mapping Table of C

<table>
<thead>
<tr>
<th>Name</th>
<th>Nome</th>
<th>Name</th>
<th>CompanyName</th>
</tr>
</thead>
<tbody>
<tr>
<td>Address</td>
<td>Indirizzo</td>
<td>Address</td>
<td>Address</td>
</tr>
<tr>
<td>Description</td>
<td>AboutUs</td>
<td>Description</td>
<td>Description</td>
</tr>
<tr>
<td>Category</td>
<td>Categoria</td>
<td>Category</td>
<td></td>
</tr>
<tr>
<td>Phone</td>
<td>Telefono</td>
<td>Tel</td>
<td>Phone</td>
</tr>
</tbody>
</table>

MT generation:
Since “AboutUs SYN Description” is in CT, these local attributes are “fused” into the same global attribute “Description”

GS annotation:
- the name and the meaning of the class Company correspond to the name and the meaning of fibre2fashion.Company (the most general class)
Global-As-View (GAV) approach:
the GS is expressed in terms of the local schemata

- **Global-as-View (GAV) mappings:**
  for each global class C we define a view $V_C$ over the local classes of C.

- The integration designer, supported by the Ontology Builder graphical interface, can implicitly define $V_C$ by the Mapping Table refinement:

  1. **Data Transformation**: converting data from local source data formats into a global schema format (Conversion Functions)
  2. **Data Fusion**: fusing records representing the same real-world object into a single, consistent, and clean record:
     1. **Object Identification**
     2. **Data Reconciliation**
Data Transformation: THALIA Benchmark

- THALIA: Test Harness for the Assessment of Legacy information Integration Approaches

  public available testbed and benchmark for information integration systems

  provides over 40 downloadable sources representing University course catalog from computer science around the world

  systematic classification of the different types of syntactic and semantic heterogeneities described by the twelve queries provided

- MOMIS Data Transformation can deal with all the twelve queries of the THALIA benchmark by using a simple combination of declarative translation functions and without the overhead of new code.
Q2: ‘Find all database courses that meet at 1:30pm on any given day’

**Complex Mappings**: Mapping between the Time attribute of Carnegie Mellon University and the Times attribute of University of Massachusetts.

<table>
<thead>
<tr>
<th>Course Mapping Table</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Course</strong></td>
</tr>
<tr>
<td>CourseTitle</td>
</tr>
<tr>
<td>Time</td>
</tr>
</tbody>
</table>

MDTF[Time][umb.Times] =

```sql
CASE WHEN ISNUMERIC(SUBSTRING(Times, 1, 2)) = 1
    THEN CASE WHEN CAST(SUBSTRING(Times, 1, 2) AS int) > 12
                THEN CAST(CAST(SUBSTRING(Times, 1, 2) AS integer) - 12 AS nvarchar(2))
                ELSE SUBSTRING(Times, 1, 2)
            END
    + SUBSTRING(Times, 3, 4) +
    CASE WHEN CAST(SUBSTRING(Times, 7, 2) AS int) > 12
            THEN CAST(CAST(SUBSTRING(Times, 7, 2) AS integer) - 12 AS nvarchar(3))
            ELSE SUBSTRING(Times, 7, 2)
        END
    + SUBSTRING(Times, 9, 3)
END AS Time
```
Mapping Refinement: Data Conflicts Resolution

- **Data Conflicts**: the same attribute from one or more sources do not agree on its value

1) **Uncertainty**: it is a conflict between a not-null value and one or more null values that describe the same attribute of the same object

2) **Contradictions**: it is a conflict between two or more different not-null values that describe the same attribute of the same object.

- **Example**: data contradictions on the **Phone** attribute

<table>
<thead>
<tr>
<th></th>
<th>L1</th>
<th>L2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Name</td>
<td>Address</td>
<td>Phone</td>
</tr>
<tr>
<td>RAMOTEX</td>
<td>…Mirpur-1216Dh</td>
<td>+390828015393</td>
</tr>
<tr>
<td>CASTORAMA</td>
<td>…Casalecchio (BO)</td>
<td>+390516113011</td>
</tr>
<tr>
<td>RAMOTEX</td>
<td>…Mirpur-1216Dh</td>
<td>880-5-801466</td>
</tr>
<tr>
<td>Koramsa Corp</td>
<td>…Guatemala City</td>
<td>+502 439 6868</td>
</tr>
</tbody>
</table>
What operator for Data Fusion?

Full Join Merge Operator

- **Full Join**: to include into the result *all tuples of all local sources*
  - Computed on the basis of the Object Identification/Join Conditions
- **Merge**: to perform data reconciliations
  - Application of Resolution functions (including all the results)

In MOMIS the **Full Join Merge** is the *default* operator, i.e., is *implicitly defined* by using the Ontology Builder graphical interface (see next slide)

The designer can change this default operator to other join operators (inner join, left/right join)
From the Mapping Table to the Full Join Merge

Mapping Table of the Global Class Hotel = \{resort, hotel\}

**Join Conditions**
- Object identifier

**Resolution Functions**
- **SUM**
  - resort
  - hotel
- **AVG**
  - Name: name, name
  - Room: rooms, hotelrooms
  - Price: amount, price
  - Star: star
  - Wifi: wifi

Select Name,
\[\text{AVG}(L1.\text{amount}, L2.\text{price})\] as Price,
\[\text{SUM}(L1.\text{rooms}, L2.\text{hotelrooms})\] as Room
...

from resort L1 **full join** hotel L2
using (name)
Data Integration and Data Fusion: an example

Global Class $G = \{L1, L2\}$

<table>
<thead>
<tr>
<th>$G$</th>
<th>$L1$</th>
<th>$L2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>ID</td>
<td>ID</td>
<td>ID</td>
</tr>
<tr>
<td>A</td>
<td>A</td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>B</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>C</td>
<td>C</td>
</tr>
</tbody>
</table>

Data Fusion

$$G \text{ as Full Join Merge of L1 and L2}$$

$$\text{SELECT ID,}$$
$$\text{L1.A AS A,}$$
$$\text{L2.B AS B,}$$
$$\text{AVG (L1.C,L2.C) AS C}$$

FROM $L1$ FULL JOIN $L2$
USING (ID)

Result

<table>
<thead>
<tr>
<th>ID</th>
<th>A</th>
<th>B</th>
<th>$C=AVG(L1.C,L2.C)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>8</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
The querying problem:
How to answer queries expressed on the GS (global queries)?

In a Virtual Data Integration system, data reside at the data sources then the query processing is based on Query rewriting: to rewrite a global query as an equivalent set of queries expressed on the local schemata data sources (local queries).

**GAV** approach: query rewriting is performed by unfolding, i.e. by expanding a global query on C according to the view associated to C.

- When the view is defined with an *outer-join merge* operator, the query rewriting performs the fusion (object identification and conflict resolution) of the local answers into the global answer.

**Query Manager**

- Distributed Query Processing
- Query Optimization
global constraint on *one-to-one attributes* can be push down on local queries

**Global Query**

**Query 1:**
```
Select Name, Room
from Hotel
where Price = 100 and Stars > 3
```

**Local queries**

Q1 to local source “resort”.

**Q1_resort:**
```
Select name, amount, rooms
from resort
where stars > 3
```

Q1 to local source “hotel”.

**Q1_hotel:**
```
Select name, price, hotelrooms
from hotel
```
The local answers \((Q_{Li})\) are fused into the global answer on the basis of the Full Outer Join-merge operation:

\[ Q_{L1} \text{ full join } Q_{L2} \text{ on } JC(L1,L2) \]

✓ **Full Join simplification:**

1. \(\text{FOJ} = Q_{L1} \text{ left join } Q_{L2} \text{ on } JC(L1, L2)\)
   - if there exists a predicate pushed down only on \(L1\)

2. \(\text{FOJ} = Q_{L1} \text{ inner join } Q_{L2} \text{ on } JC(L1, L2)\)
   - if there exists a predicate pushed down only on \(L1\) and a predicate pushed down only on \(L2\).

For Query Q1:

\[
\text{FOJ}_Q1 = \text{select } * \\
\text{from } Q1\_resort \text{ left join } Q1\_hotel \\
\text{on } Q1\_hotel .name = Q1\_resort.name
\]
**Query unfolding: Resolution Function**

**RES_FOJ**: Application of the resolution functions to FOJ

\[
\text{RES_FOJ}_\text{Q1} = \text{select} \\
\quad \text{COALESCE}(Q1\textunderscore hotel.name, \ Q1\textunderscore resort.name) \ \text{AS Name,} \\
\quad \text{SUM}(Q1\textunderscore hotel.rooms, \ Q1\textunderscore resort.hotelrooms) \ \text{AS Room,} \\
\quad \text{AVG}(Q1\textunderscore hotel.amount, \ Q1\textunderscore resort.price) \ \text{AS Price} \\
\text{from} \ \text{FOJ}_\text{Q1}
\]

**Query Result**: Application of the residual conditions to RES_FOJ

\[
\text{Query Result} = \text{select} \quad \text{Name,} \\
\quad \text{Room} \\
\text{from} \ \text{RES_FOJ}_\text{Q1} \quad \text{where Price} = 100
\]
Data Provenance for Data Integration

• **Data Provenance** or **Lineage** describes where data came from, how it was derived and how it was updated over time.

• Provenance provides valuable information that can be exploited for many purposes: managing data uncertainty, identifying and correcting data errors.

• Provenance is one of the open problems and desiderata for data fusion systems.
  - to know the origin of the visualized data
  - to explain merging decisions by tracking which original values were involved and how they have been fused.

➤ Design and Development of a Provenance Management component for the MOMIS System
Data Provenance for the MOMIS system

- **Question**: Which data provenance model for the MOMIS system?
- **Classical Data Provenance models**
  - *Lineage* encodes the tuples that were used in some derivation of the query result (set of tuples)
  - *Why-provenance* encodes all the different derivations of a tuple in the query result (set of sets of tuples)
- But they are limited to UCQs (Union of Conjunctive Queries): The full join merge of MOMIS is more expressive than UCQs
  - full join + merge (using the resolution functions).
- **PI-provenance** (Perm Influence)
  - A why-provenance model including the full join.
  - An open source system (PERM - Provenance Extension of the Relational Model) supports provenance for SQL queries.
• **PI-Provenance as a set of witness lists:**
  
  a **witness list** contains a local tuple from each local class or the special value ⊥, indicating that no tuple from a local class was used to derive the output tuple.

  ```
  L1
  ID | A  
  1  | 3  
  2  | 3  
  3  | 12 
  4  | 12 
  L2
  ID | B  
  1  | 4  
  2  |    
  3  | 3  
  ```

  ```
  SELECT PROVENANCE DISTINCT A
  FROM L1 FULL JOIN L2
  USING (ID)
  WHERE L1.A > 3 OR L2.B > 3
  ```

<table>
<thead>
<tr>
<th>A</th>
<th>PI-Provenance</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>{&lt;L1₁, L2₁&gt;}</td>
</tr>
<tr>
<td>12</td>
<td>{&lt;L1³, L2³&gt;, ⊥}</td>
</tr>
</tbody>
</table>
  ```

  i-esima tupla of L1

  L1'
**Problem:** To encode all different derivations of a tuple in the query result in presence of full join + resolution functions

**Solution:** The PI-provenance was extended to resolution functions in order to obtain all possible derivations

### A | PI-Provenance
---|---
3 | \{<L1^1, L2^1>\}
12 | \{<L1^3, L2^3>, <L1^4, ⊥>\}

```sql
SELECT DISTINCT COALESCE(L1.A, L2.A)
FROM L1 FULL JOIN L2
USING (ID)
WHERE L1.A > 3 OR L2.A > 3
```
Data Provenance in the MOMIS system: Architecture

- The "PI-provenance" is fully implemented in the "Perm" system, an open-source provenance management system.

- The "Perm" system used as the SQL engine of MOMIS (Provenance computation for the full join)

- Extensions implementation for Resolution Functions (Provenance computation for the full join-merge)
Example: Integration of biological data

- Use of the MOMIS for an integrated and unified access to many available sources of genomic and phenotypic data

CEREALAB DB stores genotypic and phenotypic cereal data collected within the CEREALAB project and integrates them with already existing data sources.
Example: Integration of the local schemata

- Two local classes (GermplasmaA and GermplasmaB) with the same attributes

  | GPN          | GermPlasm Name. |
  | FHB          | Fusarium Head Blight. |
  | Yield        | Production in t/ha. |
  | Type         | Germplasm type. |

- Global class GERMLAS (GPN,YIELD,FHB,TYPP) where
  - GPN is the shared identifier
  - YIELD,FHB and TYPE are conflicting attributes
### Example: Data Fusion

#### GPA (Germplasm A)

<table>
<thead>
<tr>
<th>GPN</th>
<th>Yield</th>
<th>FHB</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eureka</td>
<td>18</td>
<td>MR</td>
<td></td>
</tr>
<tr>
<td>Fortuna</td>
<td>7</td>
<td>MR</td>
<td></td>
</tr>
<tr>
<td>Mentana</td>
<td></td>
<td>S</td>
<td>Line</td>
</tr>
<tr>
<td>Kenora</td>
<td>20</td>
<td>MR</td>
<td>Landrace</td>
</tr>
<tr>
<td>Oasis</td>
<td>21</td>
<td>MR</td>
<td>Cultivar</td>
</tr>
</tbody>
</table>

#### GPB (Germplasm B)

<table>
<thead>
<tr>
<th>GPN</th>
<th>Yield</th>
<th>FHB</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eureka</td>
<td>6</td>
<td>S</td>
<td>Cultivar</td>
</tr>
<tr>
<td>Fortuna</td>
<td>15</td>
<td>S</td>
<td>Landrace</td>
</tr>
<tr>
<td>Mentana</td>
<td>20</td>
<td>MR</td>
<td>Line</td>
</tr>
<tr>
<td>Kenora</td>
<td></td>
<td></td>
<td>Cultivar</td>
</tr>
</tbody>
</table>

#### SQL Query

```sql
SELECT GPA.GPN,
      AVG(GPA.Yield, GPB.Yield) AS Yield,
      COALESCE(GPA.FHB, GPB.FHB) AS FHB,
      ALLVALUES(GPA.type, GPB.type) AS Type
FROM GPA
     FULL JOIN GPB
     USING (GPN)
```

<table>
<thead>
<tr>
<th>GPN</th>
<th>Yield</th>
<th>FHB</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eureka</td>
<td>12</td>
<td>MR</td>
<td>Cultivar</td>
</tr>
<tr>
<td>Fortuna</td>
<td>11</td>
<td>MR</td>
<td>Landrace</td>
</tr>
<tr>
<td>Mentana</td>
<td>20</td>
<td>S</td>
<td>Line</td>
</tr>
<tr>
<td>Kenora</td>
<td>20</td>
<td>MR</td>
<td>Landrace, Cultivar</td>
</tr>
<tr>
<td>Oasis</td>
<td>21</td>
<td>MR</td>
<td>Cultivar</td>
</tr>
</tbody>
</table>
Example: Provenance

Query: types of varieties that are resistant to FHB?

GERmplasm

<table>
<thead>
<tr>
<th>GPN</th>
<th>yield</th>
<th>FHB</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eureka</td>
<td>12</td>
<td>MR</td>
<td>Cultivar</td>
</tr>
<tr>
<td>Fortuna</td>
<td>11</td>
<td>MR</td>
<td>Landrace</td>
</tr>
<tr>
<td>Mentana</td>
<td>20</td>
<td>S</td>
<td>Line</td>
</tr>
<tr>
<td>Kenora</td>
<td>20</td>
<td>MR</td>
<td>Landrace, Cultivar</td>
</tr>
<tr>
<td>Oasis</td>
<td>21</td>
<td>MR</td>
<td>Cultivar</td>
</tr>
</tbody>
</table>

$\text{TYPE\_MR} =$

```
SELECT DISTINCT Type
FROM GERMPLASM
WHERE FHB = 'MR'
```
In the MOMIS+PERM system witness lists are represented in a relational form:

Each witness list of an output tuple is represented by a single tuple

<table>
<thead>
<tr>
<th>Type</th>
<th>GPA.GPN</th>
<th>GPA.yield</th>
<th>GPA.FHB</th>
<th>GPA.type</th>
<th>GPB.GPN</th>
<th>GPB.yield</th>
<th>GPB.FHB</th>
<th>GPB.type</th>
</tr>
</thead>
<tbody>
<tr>
<td>landrace</td>
<td>Fortuna</td>
<td>7</td>
<td>MR</td>
<td></td>
<td>Fortuna</td>
<td>15</td>
<td>S</td>
<td>landrace</td>
</tr>
<tr>
<td>cultivar</td>
<td>Eureka</td>
<td>18</td>
<td>MR</td>
<td></td>
<td>Eureka</td>
<td>6</td>
<td>S</td>
<td>cultivar</td>
</tr>
<tr>
<td>cultivar</td>
<td>Oasis</td>
<td>21</td>
<td>MR</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>landrace,cultivar</td>
<td>Kenora</td>
<td>20</td>
<td>MR</td>
<td>landrace</td>
<td>Kenora</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
S. Bergamaschi, S. Castano e M. Vincini: **Semantic Integration of Semistructured and Structured Data Sources**, SIGMOD Record Special Issue on Semantic Interoperability in Global Information, Vol. 28, No. 1, March 1999.


I. Benetti, D. Beneventano, S. Bergamaschi, F. Guerra, M. Vincini, **An Information Integration Framework for E-Commerce**, IEEE Intelligent Systems Magazine, Jan/Feb 2002, pp. 18-25,


D. Beneventano, S. Bergamaschi, F. Guerra, M. Vincini: **Building a Tourism Information Provider with the MOMIS System**, Information Technology & Tourism Journal(ISSN 1098-3058), 7:3_4, 2005.
Selected Publications


S. Bergamaschi, F. Guerra, M. Vincini, **A peer-to-peer information system for the semantic web**, in proceedings of the International Workshop on Agents and Peer-to-Peer Computing, in AAMAS 2003 Melbourne, Australia, July 14, 2003


Selected Publications


S. Bergamaschi, F. Guerra, M. Orsini, C. Sartori, **Relevant values: new metadata to provide insight on attribute values at schema level**, International Conference on Enterprise Information Systems (ICEIS 2007), 12-16, June 2007, Funchal, Madeira – Portugal

D. Beneventano, S. Bergamaschi, M. Vincini, M. Orsini, R. Carlos Nana, **Query Translation on Heterogeneous Sources in MOMIS Data Transformation Systems**, VLDB 2007 - Third International Workshop on Database Interoperability (InterDB 2007), September 24, 2007 – Vienna, Austria.


S. Bergamaschi, A. Sala: **Virtual Integration of existing web databases for the genotypic selection of cereal cultivars**, The 5th International Conference on Ontologies, DataBases, and Applications of Semantics (ODBASE 2006), Montpellier, France, Oct 31 - Nov 2, 2006
S. Bergamaschi, A. Sala, **Creating and Querying an Integrated Ontology for Molecular and Phenotypic Cereals Data**, Special Session on Agricultural Metadata & Semantics of the 2nd International Conference on Metadata and Semantics Research (MTSR’07), Corfù, Greece, October 11-12, 2007


