APPENDIX

Outline

The MOMIS system

Index

- The MOMIS system
- Academic approaches for data integration
  - Description
  - Comparison
- Bibliography

The DBGROUP @UNIMORE

- The **DataBase Group** (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 (Phd - senior researcher)
  - Francesco Guerra (Phd - researcher)
  - Mirko Orsini (Phd - research collaborator)
  - Laura Po (Phd - research collaborator)
  - Antonio Sala (Phd student)
  - Serena Sorrentino (Phd student)
  - Alberto Corni (Phd - research collaborator)

DB Group research activity

- **Intelligent Database Systems**
  - to develop Intelligent Database Systems by coupling Artificial intelligence (Description Logics) and database techniques

- **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

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)
- Project Partecipation: "WINK (Web-linked Integration of Network-based Knowledge)" supported by IST-UE RDT (cluster EUTIST-AMI) (2002-2003)

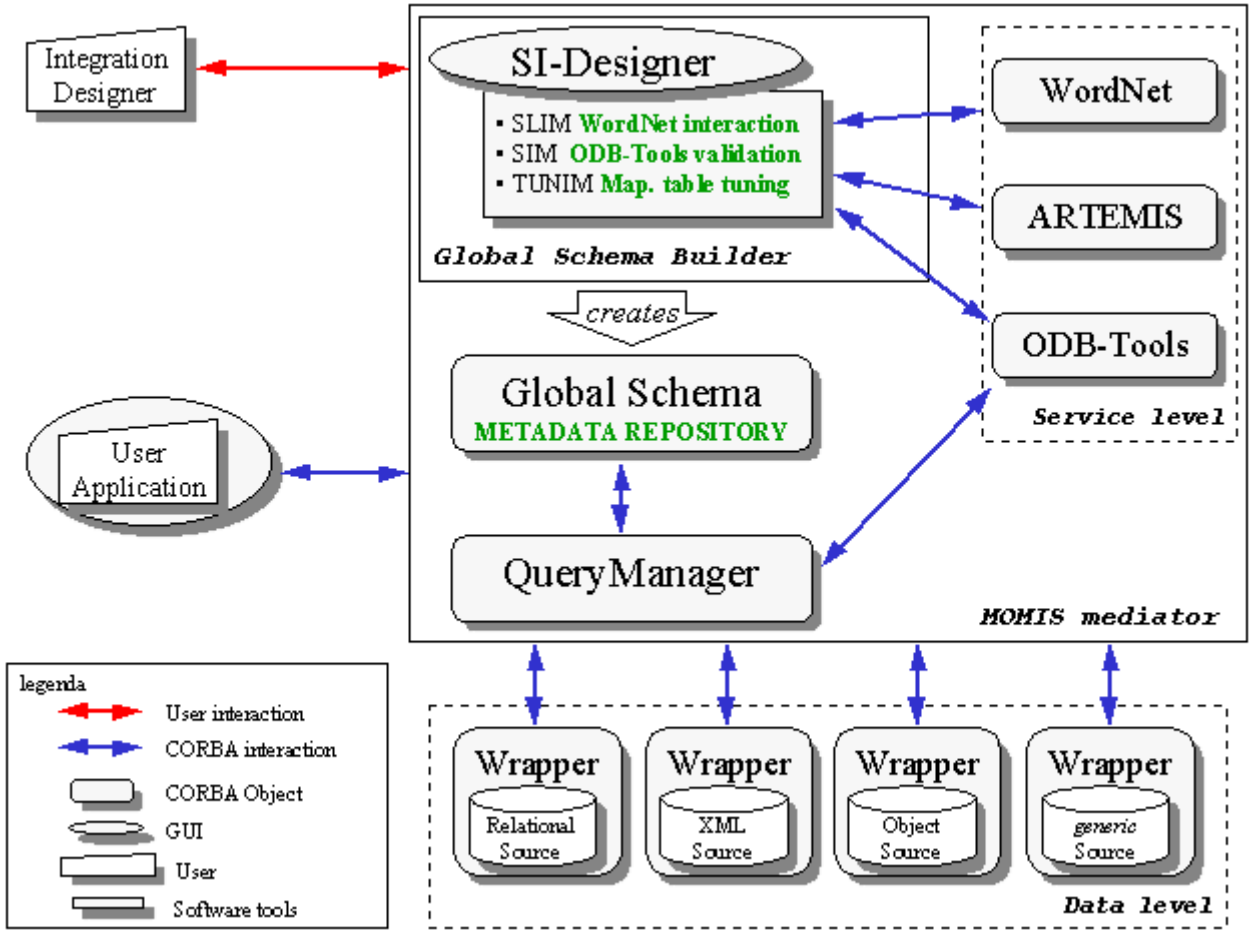
National and International Research Projects

- 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 Coordination: "**NeP4B** (Networked Peers for Business) MIUR supported by MIUR "Programma Stategico"(2006-2009)- started on July 2006
- Project Participation: "**STASIS** (SofTware for Ambient Semantic Interoperable Services)" (2006-2008) supported by IST-EU RDT - started on september 2006

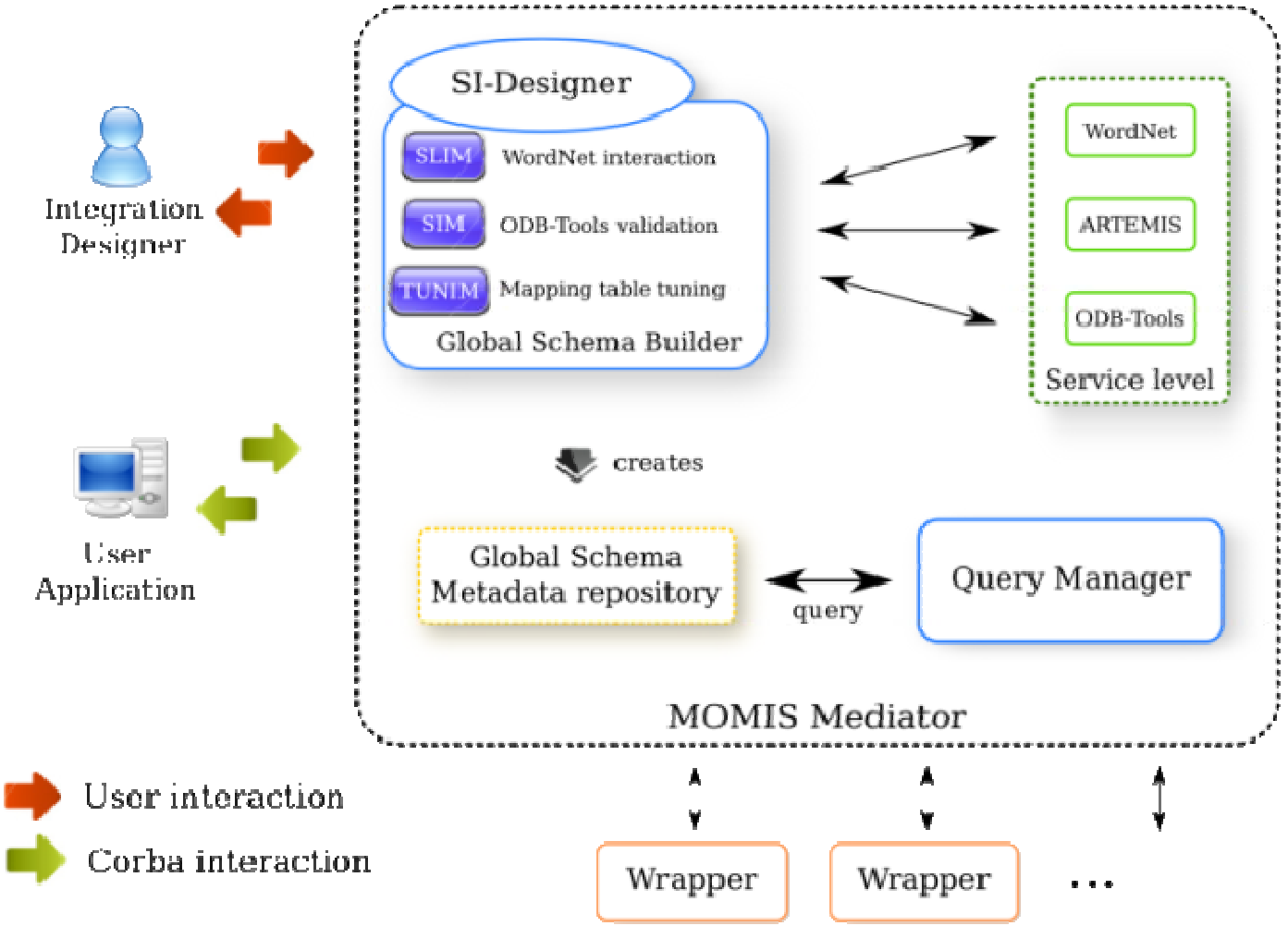
DataRiver

- **DataRiver** (www.dbgroup.unimo.it/datariver/) is a Start-Up company of the Università di Modena e Reggio Emilia started on 17 June 2009 by professors and researchers of the **DBGroup** and the **Quix** S.r.l. software house operating in the Information Technology since 10 years
- DataRiver is skilled on Data Integration, Semantic Web and Business Intelligence
- **DataRiver** designs and develops **solutions for Data Integration** solving the problems of data heterogeneity, inconsistencies and cleaning
  - Allowing to create **information value** derived by **integrated data** to improve data analysis, decision processes
- **DataRiver** developed an **Open Source Data Integration System for merging distributed and heterogeneus data**, that is able to integrate distributed information sources in a semi-automatic way following a virtual approach that preserve the autonomy and the security of the information sources

MOMIS Architecture



MOMIS Architecture



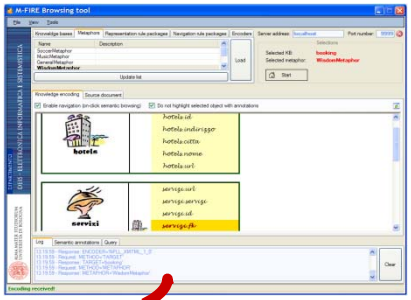
MOMIS

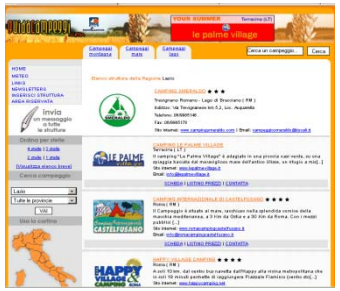
- **MOMIS** has been already tested in the above mentioned research projects for the development of Vertical Web Portals and the integration of heterogeneus data sources in many domains:
  - Tourism (vertical web portal - WISDOM)
  - Textile (search engine - SEWASIE)
  - Mechanical (search engine - SEWASIE)
  - 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

MOMIS: Tourism domain

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

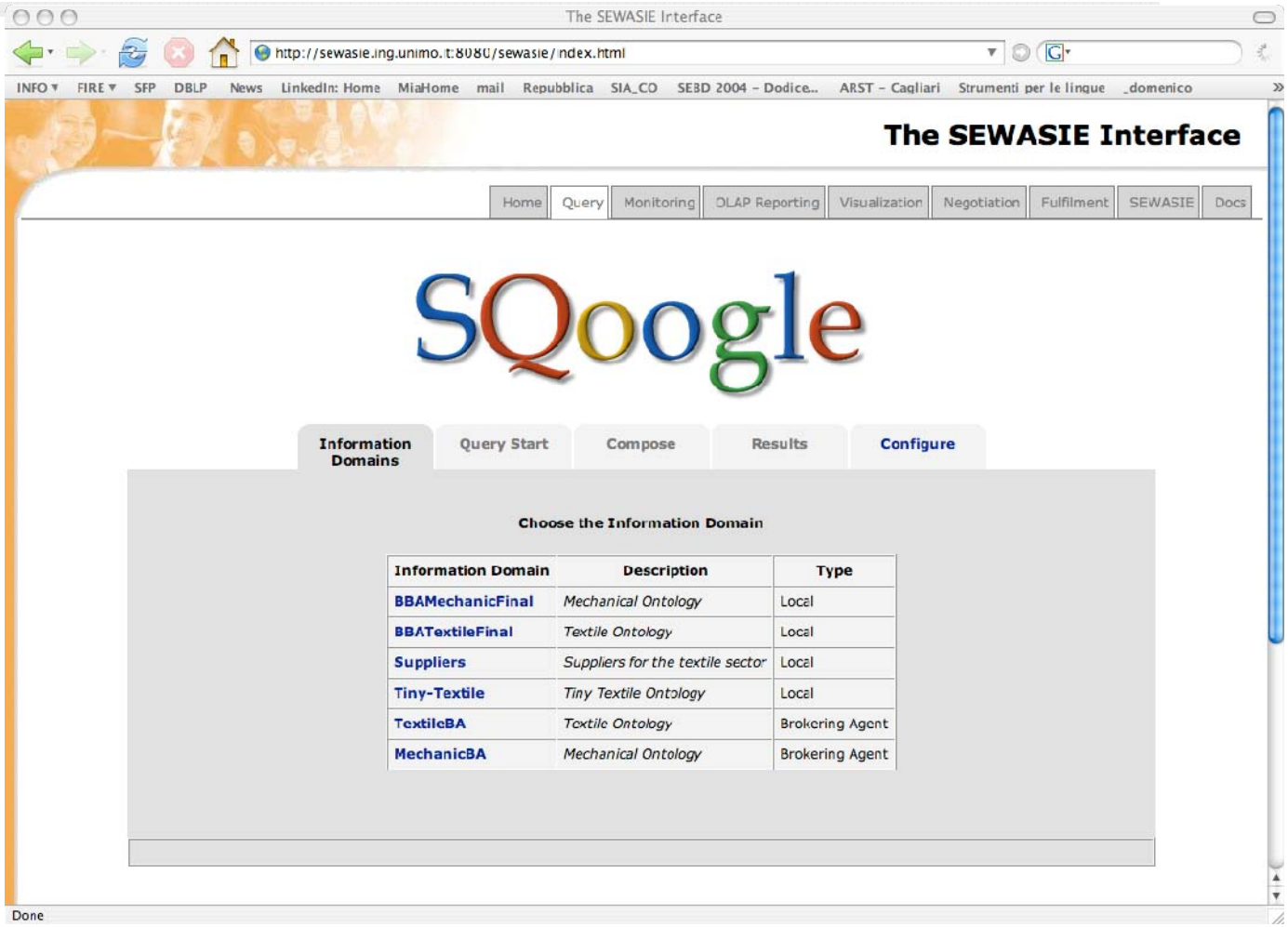




• • •

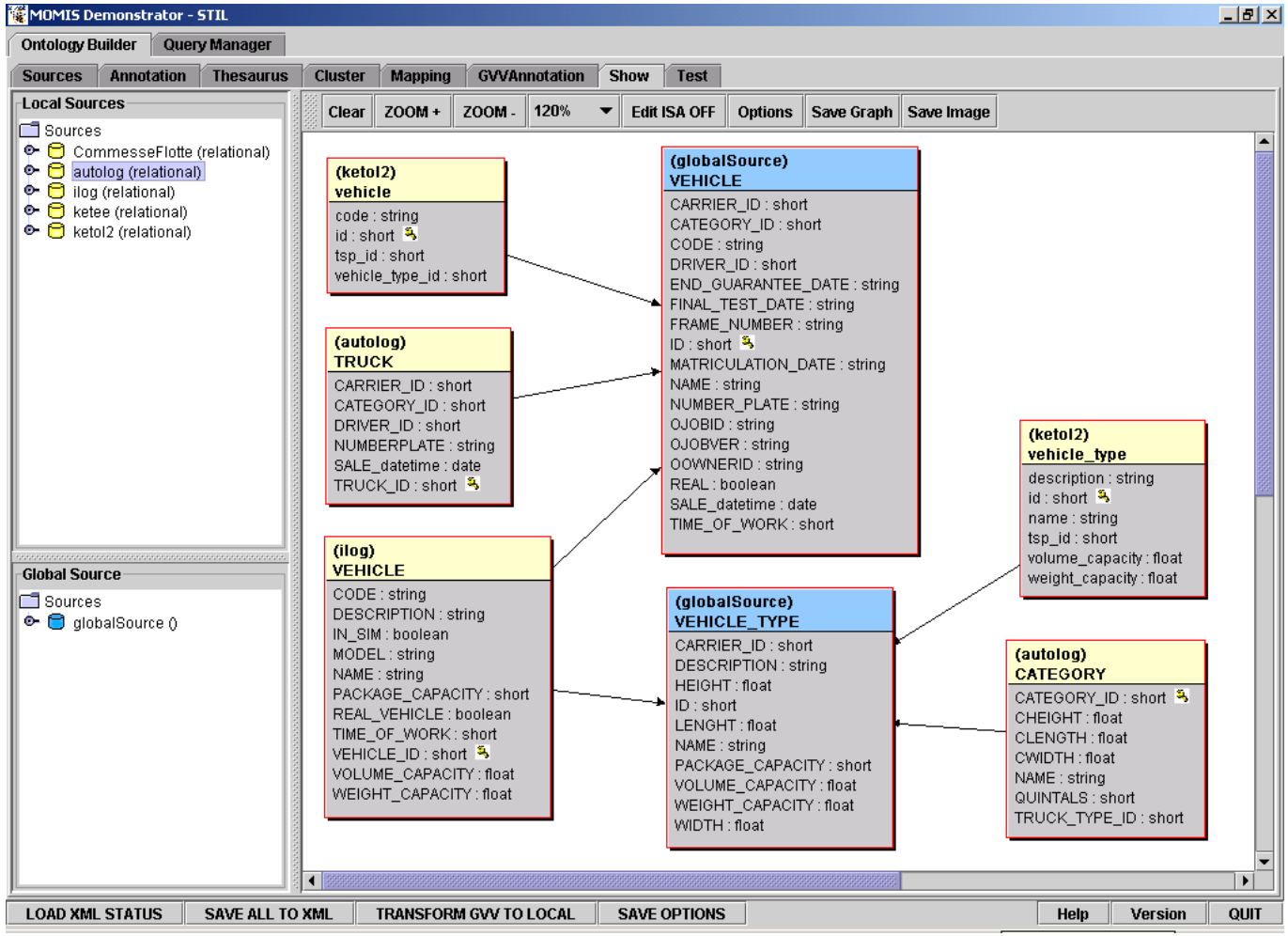
MOMIS: Textile and Mechanical domains

In the SEWASIE project (SEmantic Web and AgentS in Integrated Economies) (www.sewasie.org) the MOMIS system was exploited for the integration of heterogeneus company data sources and the development of a Semantic Search Engine

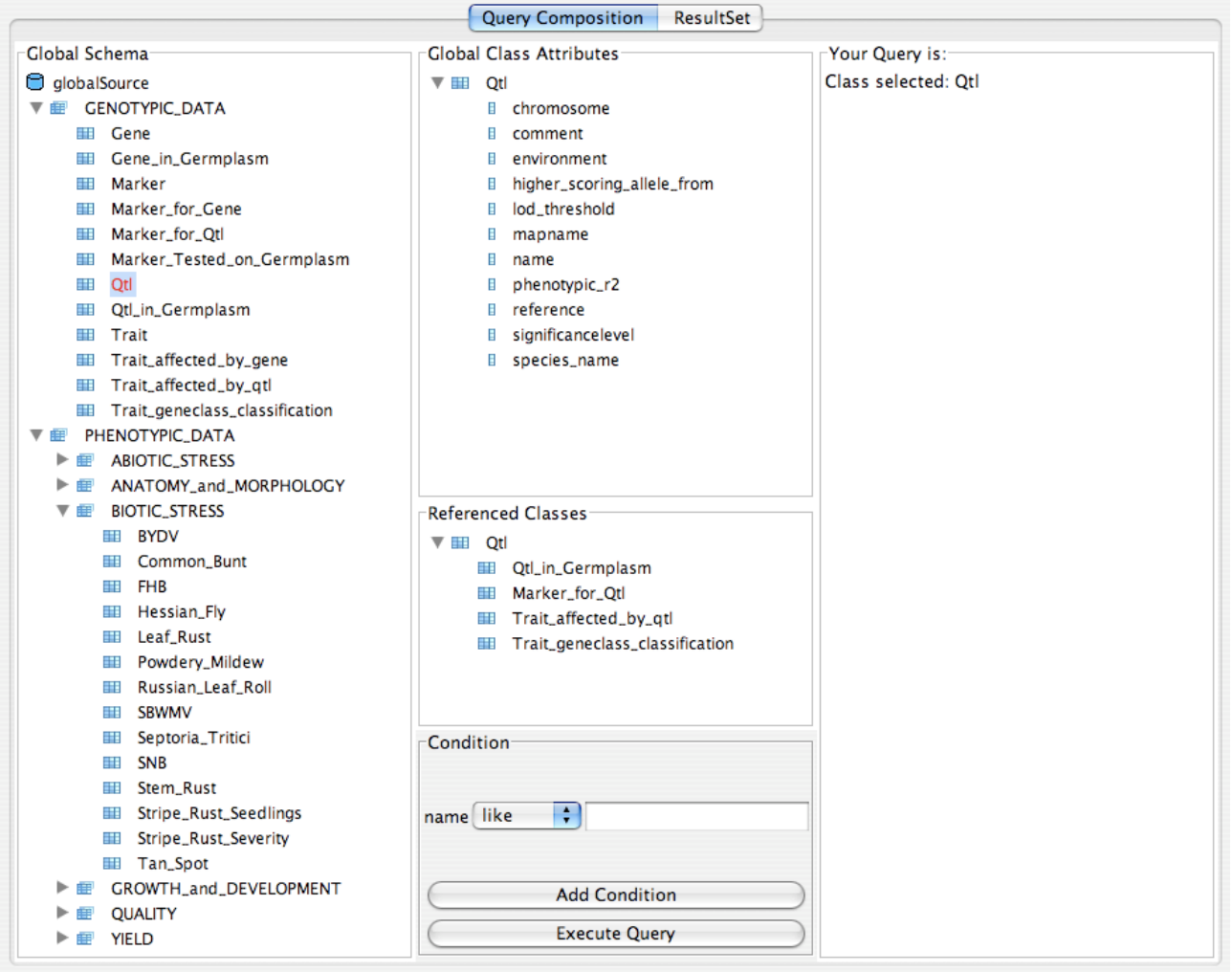


MOMIS: Logistics domain

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

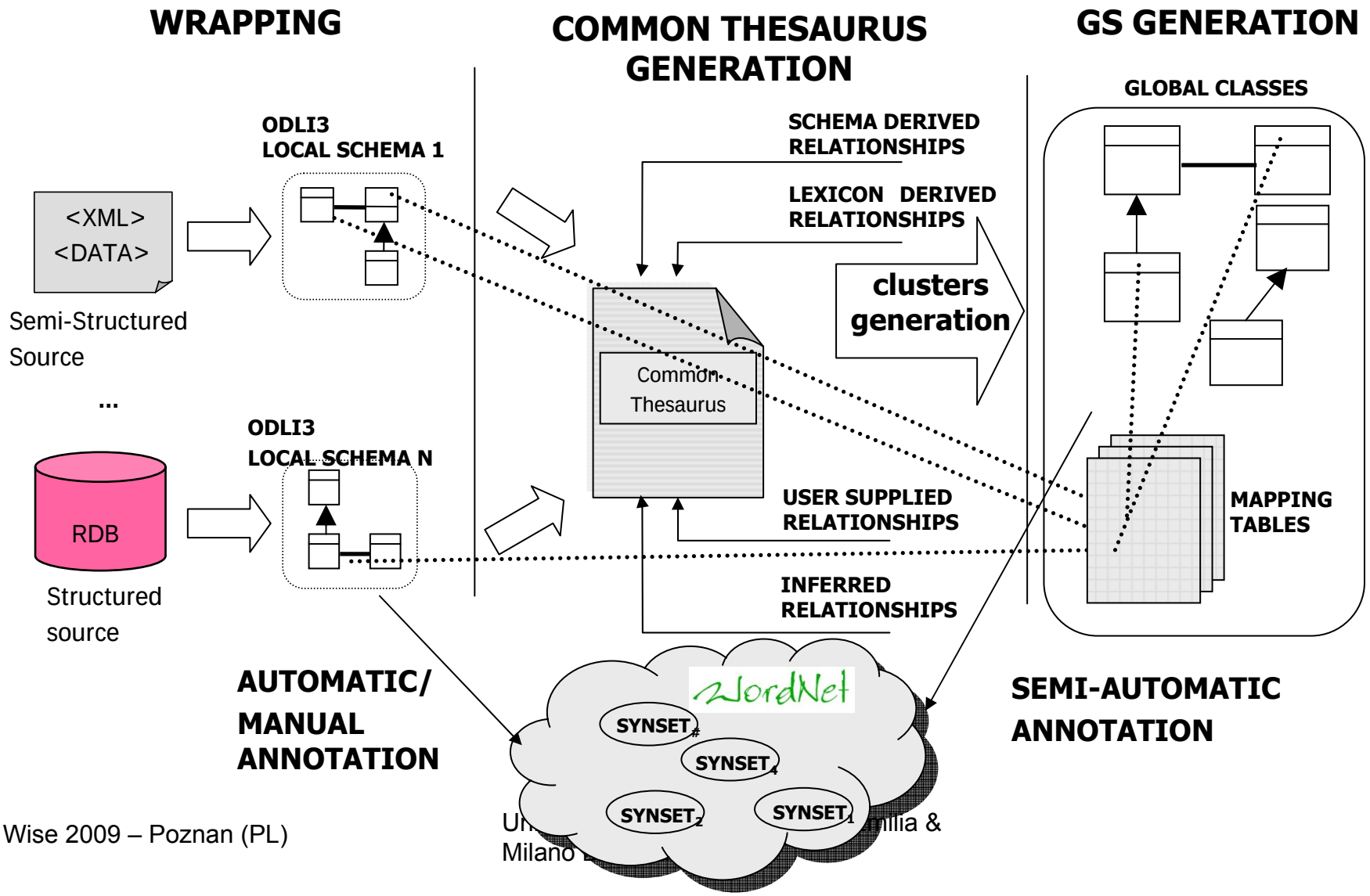


MOMIS: Agro-Food domain



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 a an integrated information system for cereals breeders

Overview of the GS generation process



Annotation and Lexicon-derived Relationships

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

NT

Hyponymy

	Word Form			
Meaning (synset)	*Regione* (region)	*Comune* (municipality)	*Description* (description)	*AboutUs* (description)
an urban district having corporate status and powers of self-government		√		
the extended spatial location of something	√			
a statement that represents something in words			√	√

Common Thesaurus relationships

prontocomune.Indirizzo.Comune	NT	prontocomune.Indirizzo.Regione
usawear.Company.Description	SYN	fibre2fashion.Company.AboutUs

Extending WordNet

- WordNet Editor [R. Benassi, S. Bergamaschi, A. Fergnani, D. Miselli: "Extending a Lexicon Ontology for Intelligent Information Integration", ECAI2004]
  - 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

- Set of intensional and extensional relationships expressing intra-schema and interschema knowledge
- **Intensional Relationships** between class and attribute name

  <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**:

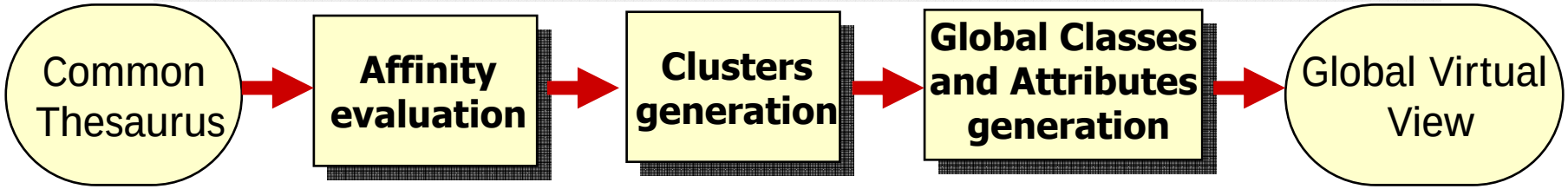
  (1) *schema derived relationships*

  (2) *lexicon derived relationships*

  (3) *designer supplied relationships*

  (4) *inferred relationships (exploiting ODB-Tools capabilities)*

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
C = {prontocomune.Azienda,fibre2fashion.Company,usawear.Company}

Mapping Table

	prontocomune.Azienda	fibre2fashion.Company	usawear.Company
Name	Nome	Name	CompanyName
Address	Indirizzo	Address	Address
Description		AboutUs	Description
Category	Categoria	Category	
Phone	Telefono	Tel	Phone

- **MT generation :**
  - Since *AboutUs SYN Description*, these local attributes correspond to the same global attribute *Description*

Mapping Table Refinement

- The integration designer, supported by the Ontology Builder graphical interface, can **define** a view associated to a global class by the **Mapping Table refinement**:

1. ***Data Transformation :*** *converting data from local source data formats into a global common 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***

Mapping Table Refinement: Data Transformation

	Name	Address	Description	Category	Phone
L1.Azienda	Nome	Indirizzo	NULL	Categoria	Telefono
L2.Company	CompanyName	Address	Description	NULL	Phone

StringConcatenation('+39 ', Telefono)

For each local class L, a **Data Transformation Operator**, T_L, is defined

L1

Nome	Indirizzo	Categoria	Telefono
RAMOTEX	…Mirpur-1216Dh	101	0828015393
CASTORAMA	…Casalecchio (BO)	586	0516113011

T_L1

Name	Address	Category	Phone
RAMOTEX	…Mirpur-1216Dh	101	+390828015393
CASTORAMA	…Casalecchio (BO)	586	+390516113011

L2

CompanyName	Address	Description	Phone
RAMOTEX	…Mirpur-1216Dh	we are dealing…	880-5-801466
Koramsa Corp	…Guatemala City	…full package	+502 439 6868

T_L2

Name	Address	Description	Phone
RAMOTEX	…Mirpur-1216Dh	we are dealing…	880-5-801466
Koramsa Corp	…Guatemala City	…full package	+502 439 6868

- Transformation of the local class instances into the GS instances is performed on the Mapping Table (by selecting conversion functions)

Mapping Table Refinement: Object Identification

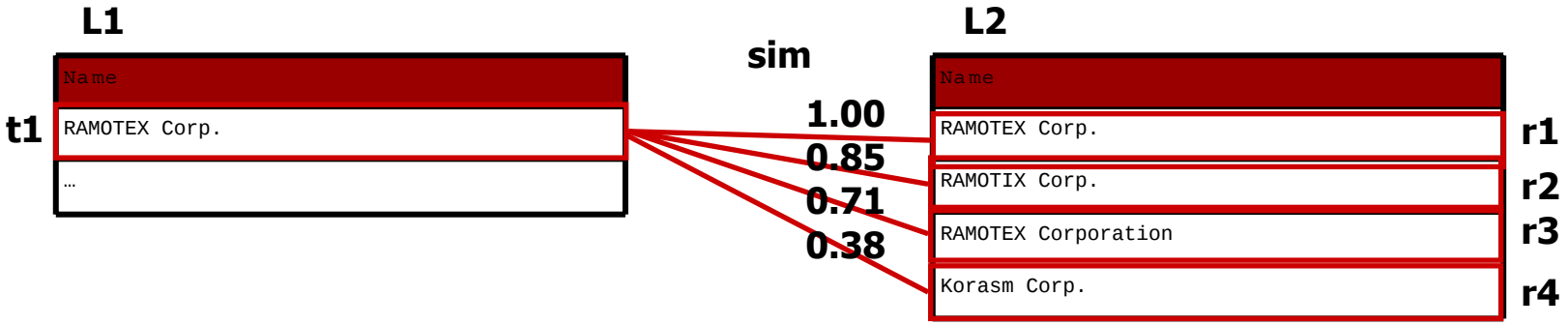
- To identify instantiation of the same object in different sources (also known as *record linkage*, *duplicate detection*, *reference reconciliation*, and many others)

- In principle **object identification** is simple: Compare each pair of objects using a similarity measure and apply a threshold. If a pair is more similar than the given threshold it is declared to be a duplicate

- In reality there are two main difficulties to be solved:
  1. effectiveness : is mostly affected by the quality of the similarity measure and the choice of a similarity threshold
  2. efficiency : data sets are often very large so even calculating and storing all pairs of objects can become an obstacle

Object Identification

Two records are similar if sim(ri, sj) ≥ θ, for some string similarity function sim() and a threshold θ



- If threshold θ = 0.7 then *t1, r1, r2, r3* will be considered the *same object* and then will be *fused together*.
- **Object Identification in MOMIS**
  - *Exact matching* (sim=1.00) expressed by **Join Condition**:

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

- We are developing a object identification method based on the sorted neighborhood method and extended to consider more than two local sources

Mapping Refinement: Data Conflicts Resolution

- **Data Conflicts** : the same attribute from one or more sources do not agree on the 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 object different not-null values that describe the same attribute of
  3. the same
- Example: data contradictions on the Phone attribute

L1

Name	Address	Phone
RAMOTEX	…Mirpur-1216Dh	+390828015393
CASTORAMA	…Casalecchio (BO)	+390516113011

Name	Address	Phone
RAMOTEX	…Mirpur-1216Dh	880-5-801466
Koramsa Corp	…Guatemala City	+502 439 6868

Data conflicts among Sources: different approaches

Name	Address	Phone
RAMOTEX	…Mirpur-1216Dh	+390828015393
CASTORAMA	…Casalecchio (BO)	+390516113011

Name	Address	Phone
RAMOTEX	…Mirpur-1216Dh	880-5-801466
Koramsa Corp	…Guatemala City	+502 439 6868

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

Name	Phone
CASTORAMA	+390516113011
Koramsa Corp	+502 439 6868

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

Name	Phone
CASTORAMA	+390516113011
RAMOTEX	+390828015393, 880-5-801466
Koramsa Corp	+502 439 6868

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

Name	Phone
CASTORAMA	+390516113011
RAMOTEX	***RF***(+390828015393, 880-5-801466)
Koramsa Corp	+502 439 6868

Mapping Refinement: 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 (*Quality-Based Resolution Functions*)
  - **Random function**
- **All Values** (Maintaining the conflicts)
- Resolution functions for **numerical attributes** : SUM, AVG, ..

Data Quality

- To improve extensional data integration techniques, data at sources can be annotated with ***data quality*** metadata
- *Data quality dimensions* supported:
  - *(syntactic) accuracy*
  - *completeness*
  - *consistency*
  - *currency*
- Quality metadata are exploited in two essential tasks
  - Object Identification
  - Quality-based resolution functions

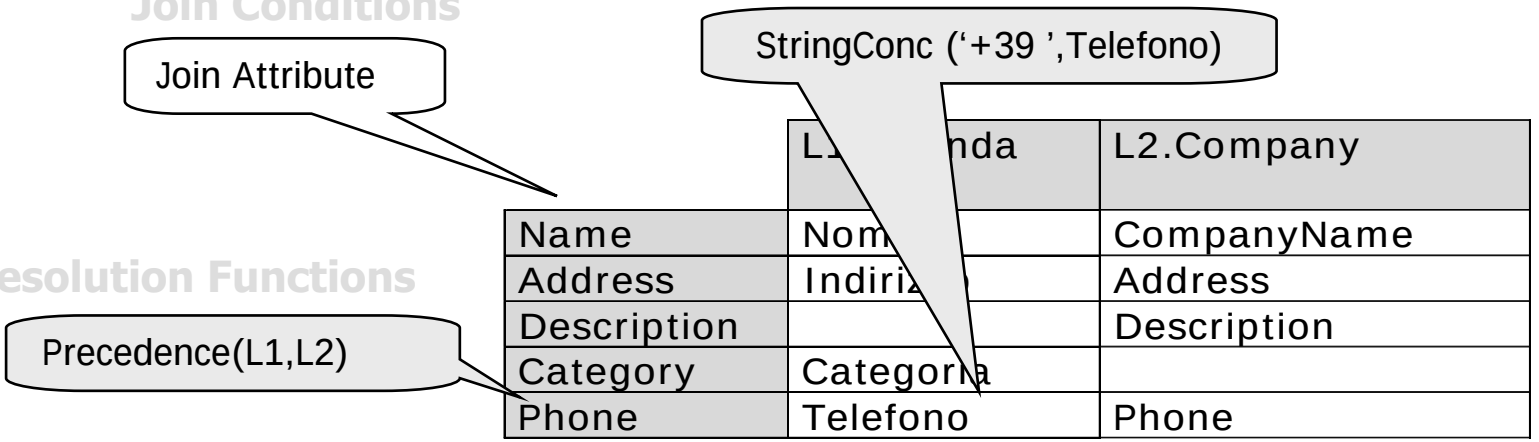
Data fusion: Outer-Join Merge

- What operator for Data Fusion ?
- **Outer-Join Merge** Operator
  - **Outer 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
- In MOMIS the Outer-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)

Building the Mappings: an example

Join Conditions

Data Conversion Functions



Resolution Functions

Precedence(L1,L2)

```
Select NAME,
       precedence(T_L1.Telefono, T_L2.Phone) as Phone,
       T_L1.Categoria,
       …
```

```
from T_L1 outer join T_L2
   on (T_L1.Nome = T_L2.CompanyName)
```

Outer Join

Outer Join Merge: an example

L1

Name	Address	Phone
RAMOTEX	…Pallabi, Mirpur-1216Dh	+390828 015393
CASTORAMA	…Casalecchio (BO)	+390516 113011

L2

Name	Address	Phone
RAMOTEX	…Pallabi, Mirpur-1216Dh	880-5-801466
Koramsa Corp	37 Ave, Guatemala City, Guatemala	+502 439 6868

L1 outer join L2 on L1.Name = L2.Name

L1.Name	L1.Address	L1.Phone	L2.Name	L2.Address	L2.Phone
CASTORAMA	…Casalecchio (BO)	+390516113011			
RAMOTEX	…Pallabi, Mirpur-1216Dh	+390828015393	RAMOTEX	…Pallabi, Mirpur-1216Dh	880-5-801466
			Koramsa Corp	37 Ave, Guatemala City, Guatemala	+502 439 6868

Application of the resolution functions

Name	Address	Phone
CASTORAMA	…Casalecchio (BO)	+390516113011
RAMOTEX	…Pallabi, Mirpur-1216Dh	+390828015393
Koramsa Corp	37 Ave, Guatemala City, Guatemala	+502 439 6868

Global Query Management

- **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

Query unfolding example

Global Class: **Company** = { **L1.Azienda**, **L2.Company**}

Global Query

```
Q:
SELECT A.Name, A.Address, A.Description, A.Category, A.Phone
FROM Company as A
WHERE A.Name LIKE '%RAM%'
AND A.Category > 100
```

Local queries

```
LQ1
SELECT Azienda.Nome, Azienda.Indirizzo, Azienda.Categoria,
Azienda.Telefono
FROM Azienda
WHERE (Nome LIKE '%RAM%') AND (Categoria > 100)
```

```
LQ2
SELECT Company.CompanyName, Company.Address,
Company.Description, Company.Phone
FROM Company
WHERE CompanyName LIKE '%RAM%'
```

Query unfolding example

```
Q:
SELECT A.Name, A.Address, A.Description, A.Category, A.Phone
FROM Company as A
WHERE A.Name LIKE '%RAM%'
AND A.Category > 100
```

```
Outer Join computation:
OJ :     SELECT        *
         FROM T_LQ1 full outer join T_LQ2
                on (T_LQ1.Nome = T_LQ2.Name)
```

```
Residual predicate:
RP:      SELECT  *
         FROM OJ
         WHERE Category > 100
```

```
Query Answer
Select  Name, Address, Description, Category,
         precedence(T_LQ1.Telefono, T_LQ2.Phone) AS Phone
from RP
```

MOMIS - references

HOME PAGE: http://www.dbgroup.unimo.it/Momis/

- 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.
- D. Beneventano, S. Bergamaschi, S. Castano, A. Corni, R. Guidetti, G. Malvezzi, M. Melchiori e M. Vincini: Information Integration: the MOMIS Project Demonstration, International Conference on Very Large Data Bases (VLDB'2000), Cairo, Egypt, Settembre 2000.
- S. Bergamaschi, S. Castano, D. Beneventano e M. Vincini: Semantic Integration of Heterogeneous Information Sources, Special Issue on Intelligent Information Integration, Data & Knowledge Engineering, Vol. 36, Num. 1, Pages 215-249, Elsevier Science B.V. 2001.
- D. Beneventano, S. Bergamaschi, F. Guerra, M. Vincini: Synthesizing an Integrated Ontology, IEEE Internet Computing, Vol.7,N.5, September/October 2003.
- 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.

Developed systems

Academic approaches

Academic approaches

- It is not possible to build an exhaustive list of approaches for data integration developed by the research community
  - Several important systems were developed between 1995 – 2002
  - Later on the research moved on:
    - The application of similar techniques for ontology management in the context of the semantic web
    - Developing techniques and architectures improving the process automation (integration on the fly) dealing also with uncertainty, probability
    - Entity recognition / record linkage
    - Keyword based search engine

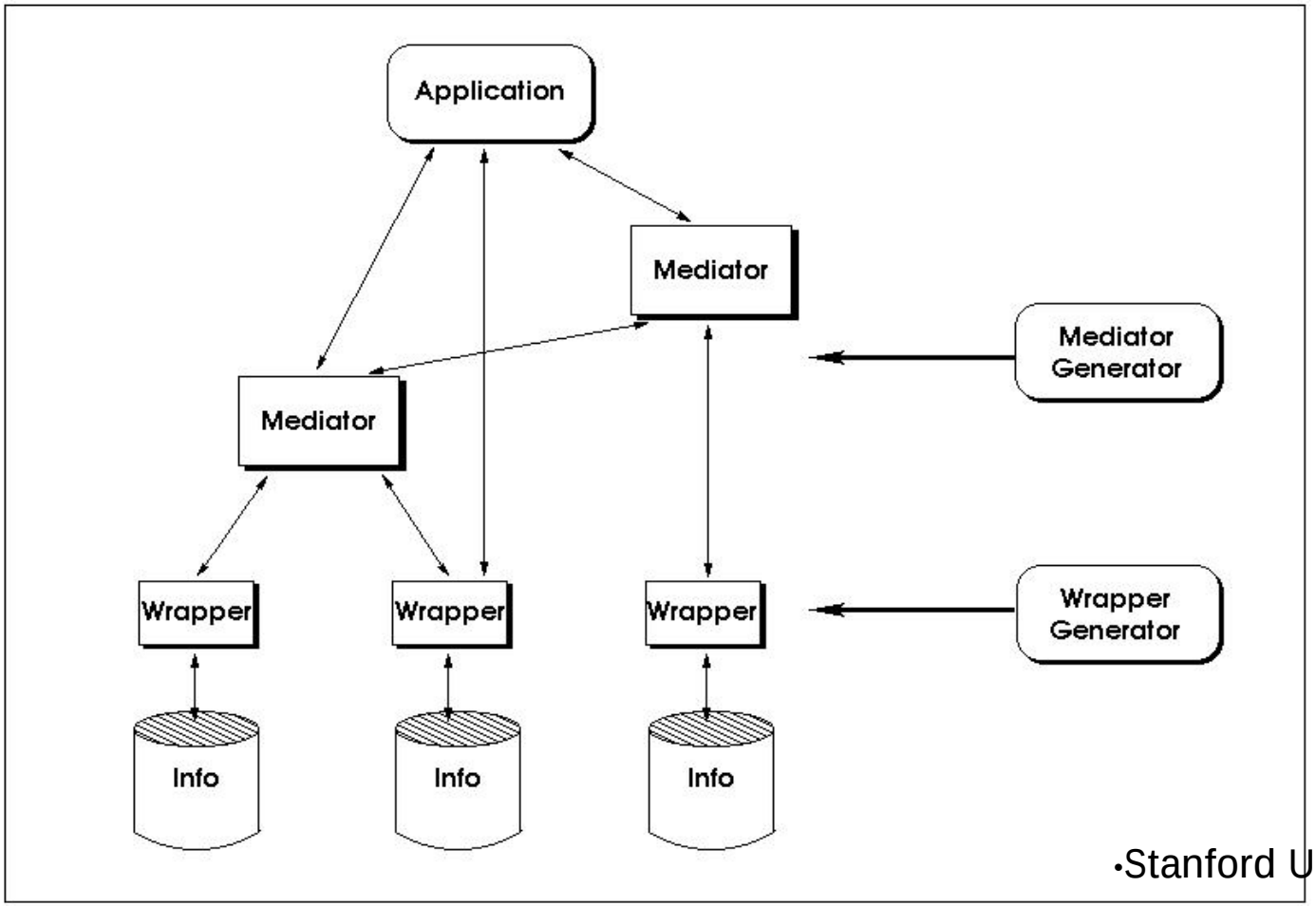
The academic systems pioneers in data integration

- TSIMMIS
- GARLIC
- SIMS
- Information Manifold
- Infomaster
- COIN
- MOMIS

TSIMMIS

- The TSIMMIS system (University of Stanford) [Li 1998] is definitely one of the most well-known system for data integration
- TSIMMIS integrates data from multiple heterogeneous sources and provides users with seamless integrated views of the data
- It translates a user query on the integrated views into a set of source queries and post processing steps that compute the answer to the user query from the results of the source query
- One of the distinguishing features of TSIMMIS is its use of a semi-structured data model(called the Object Exchange Model or OEM) for dealing with the heterogeneity of the data sources

TSIMMIS – 3 layer architecture



- Stanford University
- www-db.stanford.edu/tsimmis

TSIMMIS OEM: Object Exchange Model

- OEM : Self describing semi-structured data model (structural approach)
  - Each object is identified by an id followed by a nested description, only base types, no inheritance

```
<id: label, type, value>
```

Example

```
<ob1: person, set, {sub1,sub2,sub3,sub4,sub5}>
   <sub1: last_name, str, 'Smith'>
   <sub2: first_name, str, 'John'>
   <sub3: role, str, 'faculty'>
   <sub4: department, str, 'cs'>
   <sub5: telephone, str, '45762345'>
```

Semi-structured: another object with the same label (person)may have different sub-objects (see XML).

MSL: Mediator Specification Language

- Rule Definition and Query Template
  - The head defines the global schema concept, the tail defines the related concepts in the data sources

```
(MS1) Rule:
<cs_person {<name N> <rel R> Rest1 Rest2}>
   :- <person {<name N> <dept 'cs'> <rel R> | Rest1
   }>@source1
   AND decomp(N,LN,FN)
   AND <R {<first_name FN> <last _name LN> |
   Rest2}>@source2
(QT1) Query ::= *O :- <O cs_person {<name $N>}>
```

TSIMMIS Main references

HOME PAGE: http://www-db.stanford.edu/tsimmis/tsimmis.html

- J. Hammer, H. Garcia-Molina, K. Ireland, Y. Papakonstantinou, J. Ullman, and J. Widom. "Information Translation, Mediation, and Mosaic-Based Browsing in the TSIMMIS System". Proc. of the ACM SIGMOD Int. Conf. on Management of Data, pp 483, 1995.
- H. Garcia-Molina, J. Hammer, K. Ireland, Y. Papakonstantinou, J. Ullman, and Jennifer Widom. "Integrating and Accessing Heterogeneous Information Sources in TSIMMIS". In Proc. of the AAAI Symposium on Information Gathering, pp. 61-64, 1995.
- Y. Papakonstantinou, H. Garcia-Molina and J. Widom. "Object Exchange Across Heterogeneous Information Sources". IEEE ICDE, pp. 251-260, 1995.
- C. Li, R. Yerneni,V. Vassalos, H. Garcia-Molina, Y. Papakonstantinou, J. Ullman: Capability Based Mediation in TSIMMIS. ACM SIGMOD 1998: 564—566

GARLIC

- Garlic (IBM Almaden Research Center) [Carey et al 1995] is able to manage quantities of heterogeneous multimedia information
  - traditional and multimedia data are stored in a variety of existing data repositories, including databases, files, text managers, image managers, video servers and so on
  - the data is seen through a unified schema expressed in an object-oriented data model
  - The data can be queried and manipulated using an object-oriented dialect of SQL

GARLIC - architecture

GARLIC follows the typical three layer architecture of mediator systems

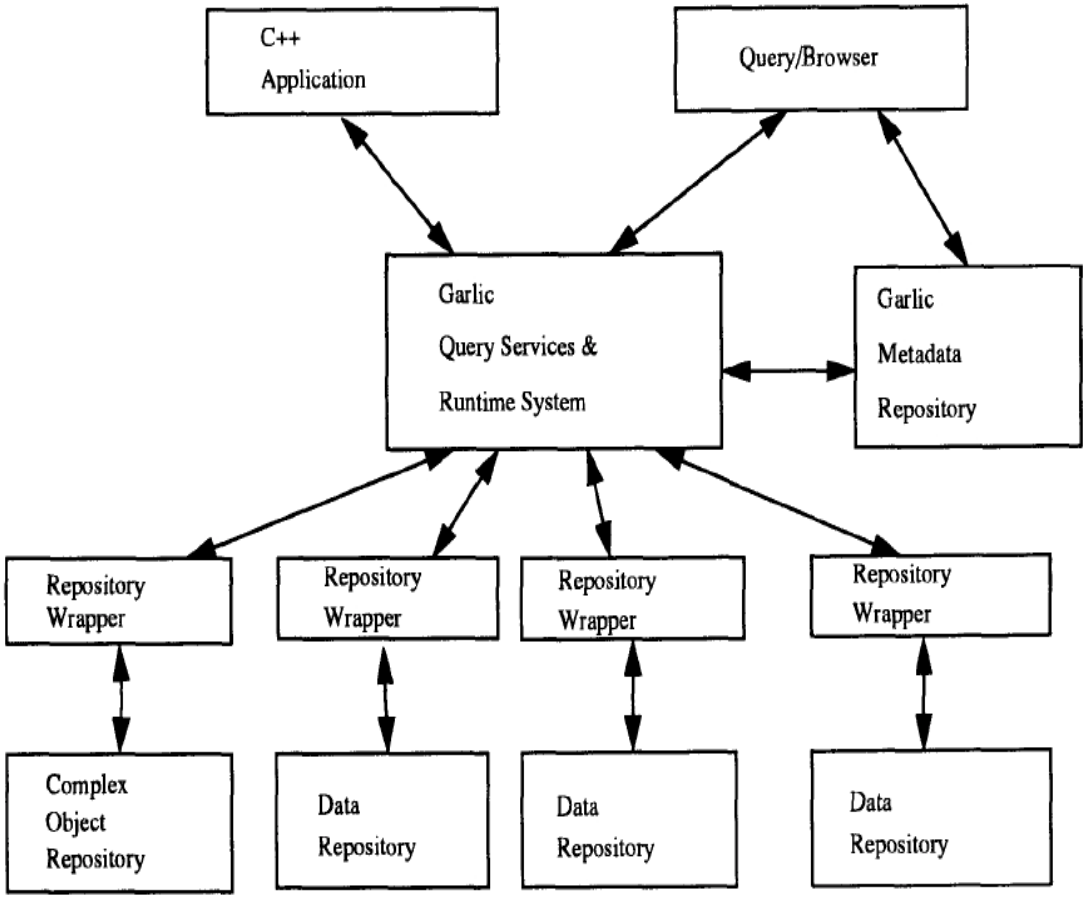


Figure 1. Garlic System Architecture

GARLIC – GDL: Garlic Data Language

Relational Repository Schema

```
interface Country {
 attribute string name;
 attribute                    string
   airlines_served;
 attribute boolean visa_required;
 attribute Image scene; }

interface City {
 attribute string name;
 attribute long population;
 attribute boolean airport;
 attribute Country country;
 attribute Image scene; }
```

Web Repository Schema

```
interface Hotel {
 attribute readonly string name;
 attribute readonly short class;
 attribute      readonly      double
   daily_rate;
 attribute readonly string location;
 attribute readonly string city; }
```

Image Server Repository Schema

```
interface Image {
 attribute      readonly      string
   file_name;
 double     matches     (in     string
   file_name);
 void     display     (in     string
   device_name); }
```

Garlic Main references

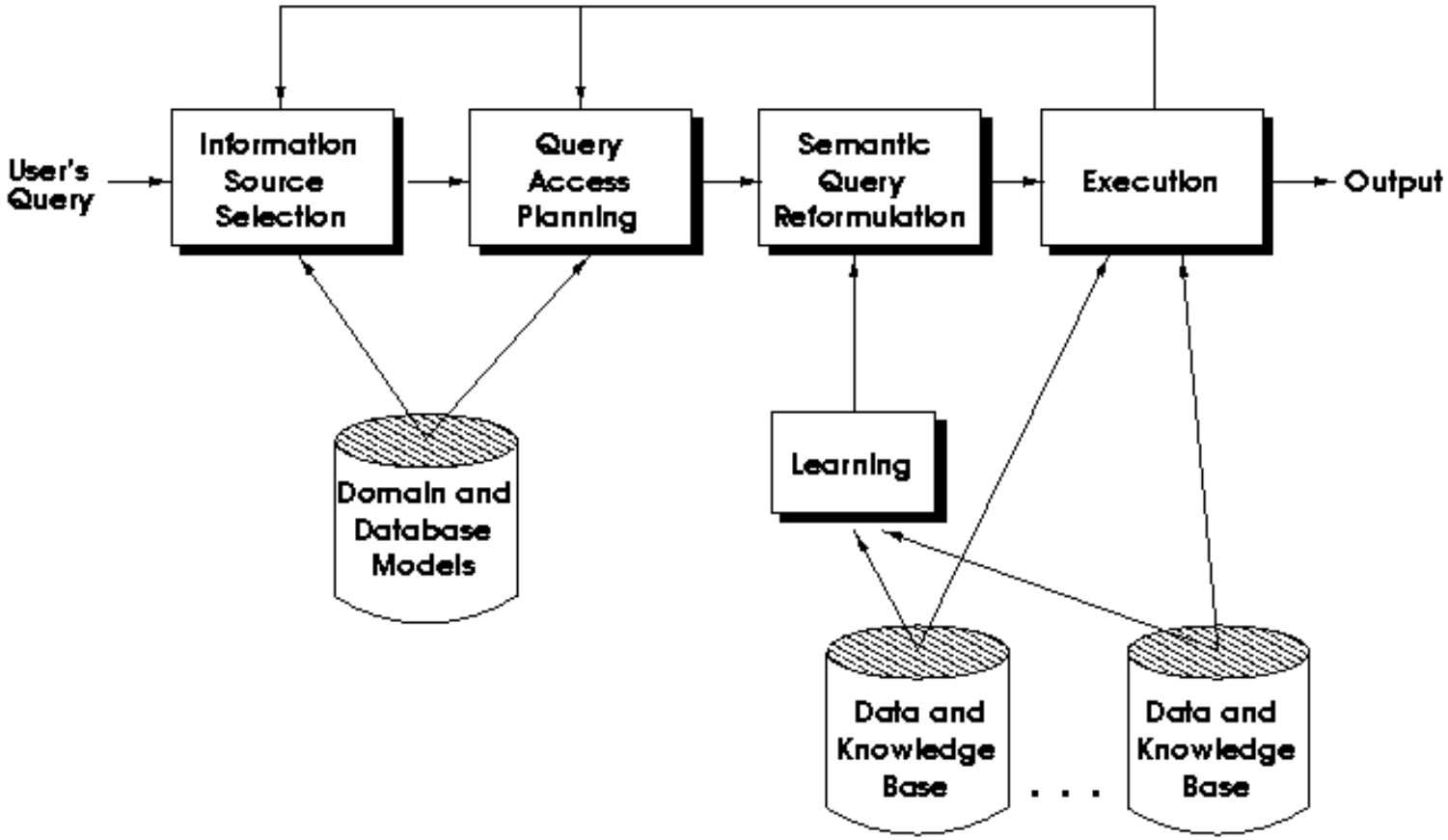
HOME PAGE:http://www.almaden.ibm.com/cs/garlic/homepage.html

- M. T. Roth, P. M. Schwarz, "Don't Scrap It, Wrap It! A Wrapper Architecture for Legacy Data Sources", Proc. of the VLDB 97, pp 266-275, 1997.
- L. M. Haas, D. Kossmann, E. L. Wimmers, J. Yang, "Optimizing Queries Across Diverse Data Sources", Proc. of the VLDB 97, pp 276-285, 1997.
- M. T. Roth, M. Arya, L. M. Haas, M. J. Carey, W. F. Cody, R. Fagin, P. M. Schwarz, J. Thomas II, E. L. Wimmers "The Garlic Project", Proc. of the1996 ACM SIGMOD, pp 557, 1996.

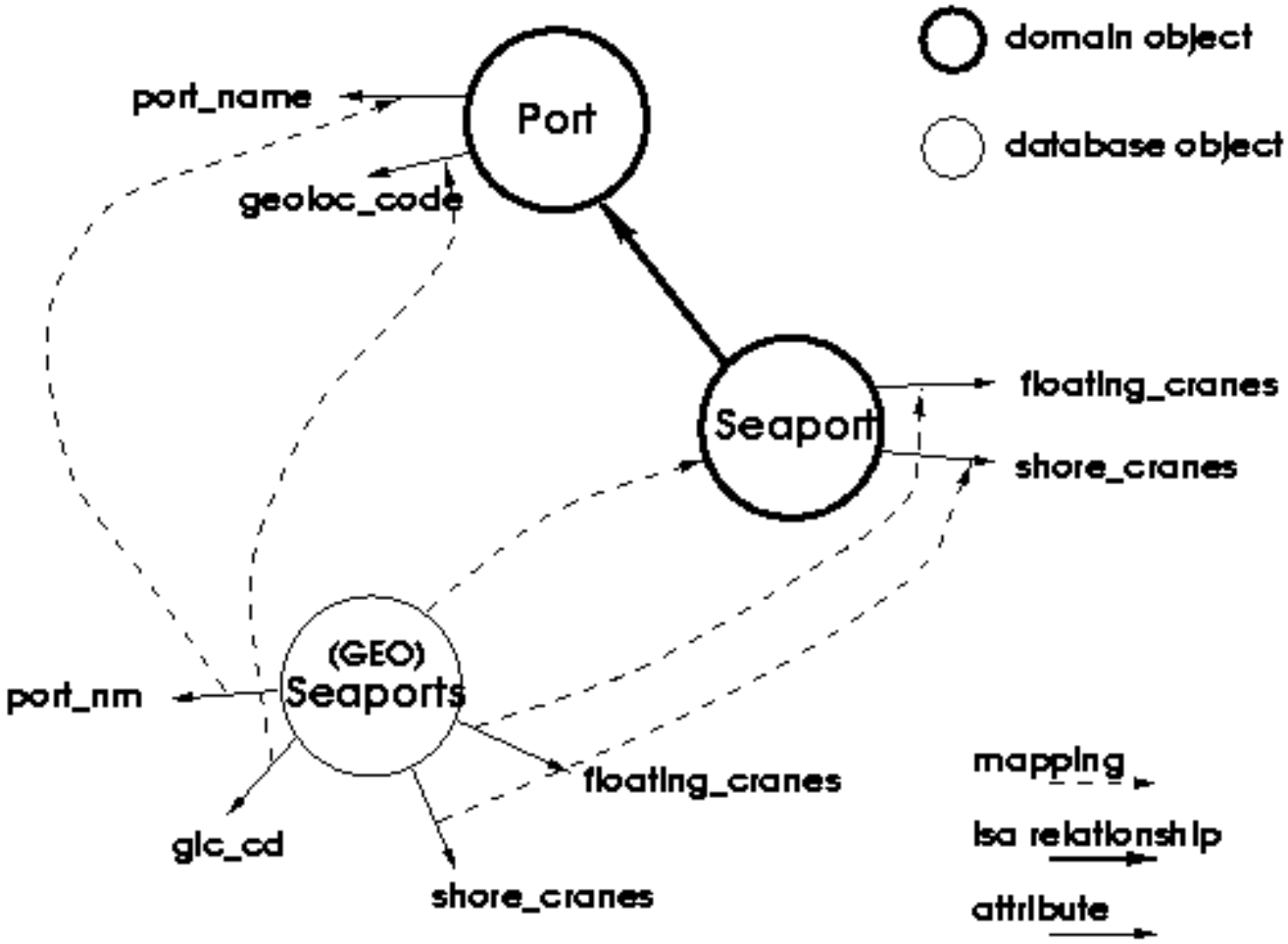
SIMS

- SIMS (Services and Information Management for decision Systems) was developed at the University of Southern California [Arens, 1997]
- Domain model: a metaschema to represent the knowledge and databases to be integrated
- LOOM Description Logics rules and semantics to define the query planning
- Algorithm to generate the query access plan:
  - select data source
  - Generalize/specialize concept
  - partition concept
  - reformulation process
  - caching retrieved data

SIMS


User's Query
Information Source Selection
Query Access Planning
Semantic Query Reformulation
Execution
Output
Domain and Database Models
Learning
Data and Knowledge Base
Data and Knowledge Base

SIMS – domain mappings



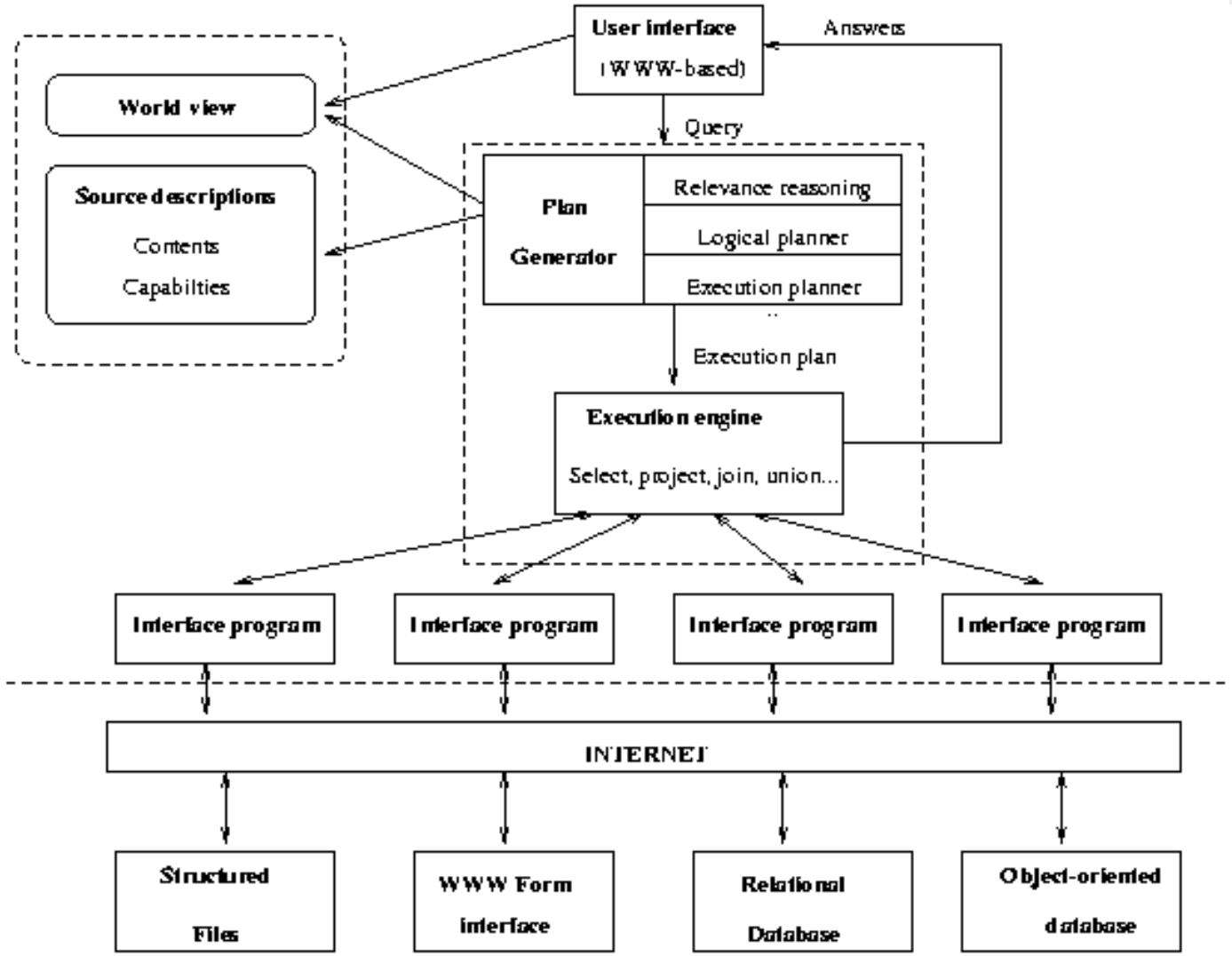
SIMS - references

- Y. Arens, C. A. Knoblock and C. Hsu "Query Processing in the SIMS Information Mediator", Advanced Planning Technology, editor, Austin Tate, AAAI Press, Menlo Park, CA, 1996.
- Y. Arens, C. Y. Chee, C. Hsu, and C. A. Knoblock "Retrieving and Integrating Data from Multiple Information Sources", Int. Journal of Intelligent and Cooperative Information Systems. Vol. 2, No. 2. Pp. 127-158, 1993.
- J. L. Ambite and C. A. Knoblock, "Reconciling Distributed Information Sources", Working Notes of the AAAI Spring Symposium on Information Gathering in Distributed Heterogeneous Environments, 1995.
- Evolution: the Ariadne project
- Naveen Ashish, Craig A. Knoblock, and Cyrus Shahabi. Selectively materializing data in mediators by analyzing user queries International Journal of Cooperative Information Systems, 11(1), March 2002.
- Craig A. Knoblock, Steven Minton, Jose Luis Ambite, Naveen Ashish, Ion Muslea, Andrew G. Philpot, and Sheila Tejada The ariadne approach to web-based information integration. International Journal of Cooperative Information Systems (IJCIS), Special Issue on Intelligent Information Agents: Theory and Applications, 10(1/2):145--169, 2001.

Information Manifold

- Information Manifold [Halevy 2006] provides a uniform query interface to a multitude of data sources, thereby freeing the casual user from having to locate data sources, interact with each one in isolation and manually combine results
- The Information Manifold proposed the method that later became known as the Local-as-View approach: an information source is described as a view expression over the mediated schema

Information Manifold



Information Manifold

- **Concept**
- **World views:** collection of virtual relation and classes of the sources (a non-materialized global schema)
- **Query:** conjunctive queries over the world view
- **Constraint** associated with the sources
- **Capabilities of sources:** the selection and join functionalities available
- **Query decomposition** (by Answering algorithms)
- **Merge of the result**
  - Based on a **Description Logics - CARIN (LAV approach)**

Information Manifold References

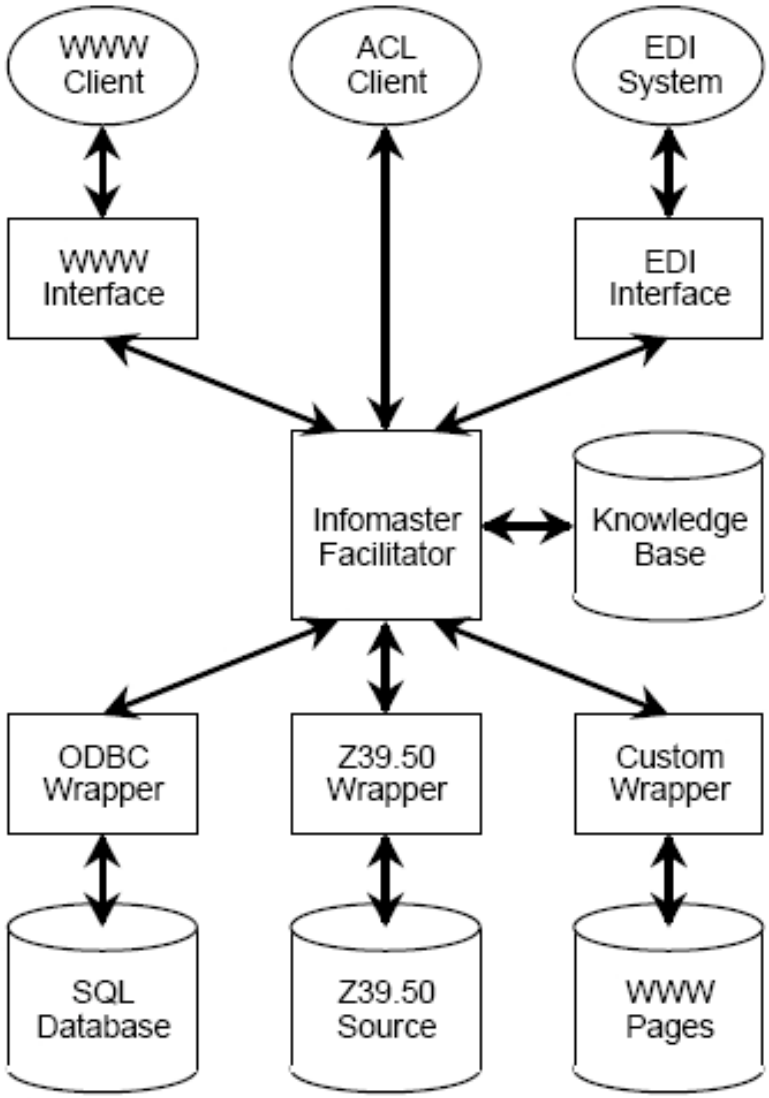
HOME PAGE: http://www.cs.washington.edu/homes/alon/

- A. Y. Levy , A. Rajaraman , J. J. Ordille, "Querying Heterogeneous Information Sources Using Source Descriptions", Proc. of the VLDB Conf. 1996.
- C. Beeri , A. Y. Levy , M. C. Rousset, "Rewriting Queries Using Views in Description Logics", Proc.of the 16th ACM SIGACT-SIGMOD-SIGART, 1997.
- A. Y. Levy, "Obtaining Complete Answers from Incomplete Databases", Proc. of the 22nd VLDB Conference, 1996.
- Alon Y. Halevy, Anand Rajaraman, Joann J. Ordille: Data Integration: The Teenage Years. VLDB 2006: 9-16
- **Nimble**:http://www.actuate.com/products/data-integration/index.asp

Infomaster

- Infomaster [Genesereth 1997] is an information integration system developed by University of Stanford,that provides integrated access to multiple distributed heterogeneous information sources
  - Infomaster creates a virtual data warehouse
  - Infomaster is a facilitator that dynamically determines an efficient way to answer the user's query
  - It provides integrated access to distributed, heterogeneous information sources, thus giving its users the desirable illusion of a centralized, homogeneous information system

Infomaster - architecture



Infomaster

- Wrappers allows the system accessing information in a variety of sources
- Infomaster uses rules and constraints to describe information sources and translations among these sources
  - These rules and constraints are stored in a knowledge.
  - For effocient access, the rules andconstraints are loaded into Epilog, a main memory database system from Epistemics
- The user interface has two levels of access: an easy-to-use, forms-based interface, and anadvanced interface that supports arbitrary constraints applied to multiple information sources
- Infomaster has a programmatic interface called Magenta, which supports ACL (Agent Communication Language) access. ACL consists of KQML (Knowledge Query and Manipulation Language), KIF (Knowledge Interchange Format), as well as vocabularies of terms

Infomaster - references

- Oliver M. Duschka, Michael R. Genesereth: Query Planning in Infomaster. SAC 1997: 109-111
- Michael R. Genesereth, Arthur M. Keller, Oliver M. Duschka: Infomaster: An Information Integration System. SIGMOD Conference 1997: 539-542
- Arthur M. Keller, Michael R. Genesereth: Using Infomaster to Create a Housewares Virtual Catalog. Electronic Markets 7(4): (1997)
- Michael R. Genesereth, Illah R. Nourbakhsh: Time-Saving Tips for Problem Solving with Incomplete Information. AAAI 1993: 724-730

COIN

- The COntext INterchange (COIN) [Goh, 1999] strategy presents a perspective for mediated data access in which semantic conflicts among heterogeneous systems are not identified a priori, but are detected and reconciled by a Context Mediator through comparison of contexts axioms corresponding to the systems engaged in data exchange
- COIN is a mediator-based approach developed by the Massachusetts Institute of Technology and the University of Singapore for achieving semantic interoperability among heterogeneous sources and receivers, constructed on the following tenets:
  - should be transparent to a user
  - the provision of such a mediation service requires only that the user furnish a logical specification of how data are interpreted in sources and receivers, and how conflicts, when detected, should be resolved, but not what conflicts exists a priori between any two systems

COIN

- The semantics of data is expressed in the form of a *context theory* and a set of *elevation axioms* with reference to a *domain model*
- .Queries are intercepted by a *Context Mediator,* which rewrites the user query to a *mediated query.*
- The *Optimizer* transforms this to an Optimized query plan, which takes into account a variety of cost information.
- The optimized query plan is executed by an *Executioner* which dispatches subqueries
- Individual systems, collates the results and undertakes conversions

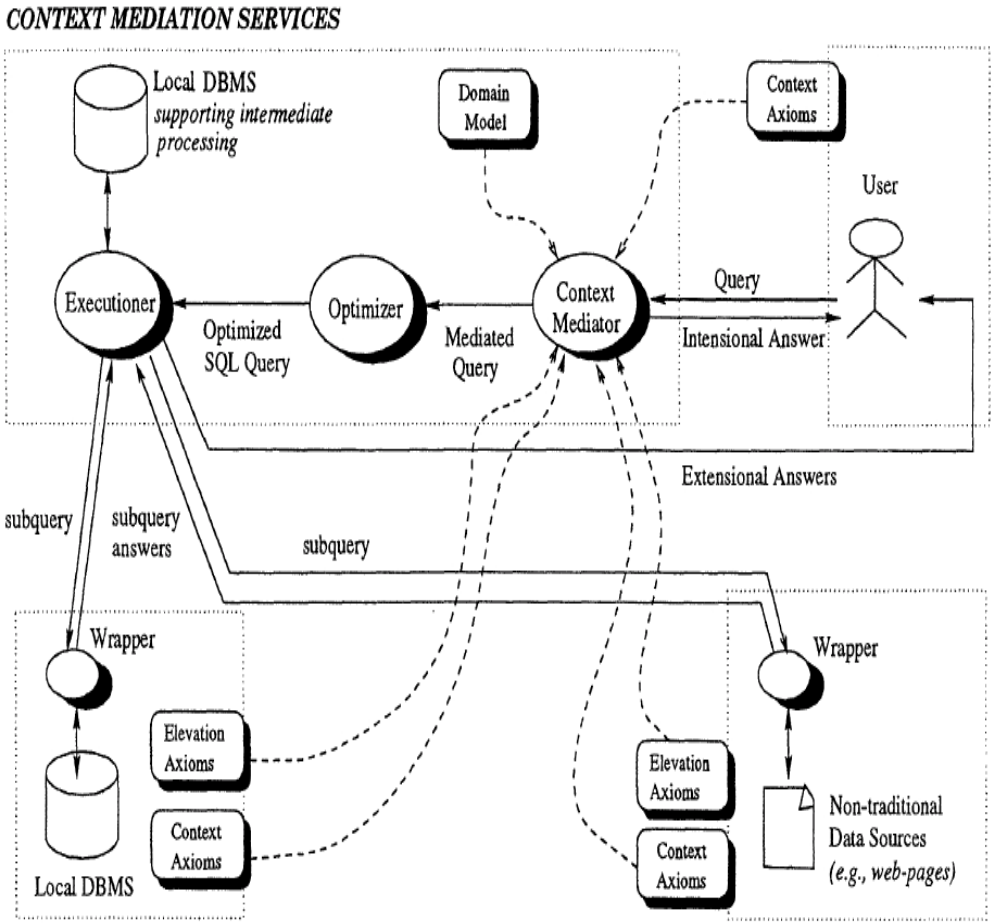


Figure 2: Architecture of a Context Interchange System

COIN - references

- Cheng Hian Goh and Stephane Bressan and Stuart Madnick and Michael Siegel: Context Interchange: New Features and Formalisms for the Intelligent Integration of Information. 1998
- Allen Moulton, Stuart E. Madnick, Michael Siegel: Semantic Interoperability in the Securities Industry: Context Interchange Mediation of Semantic Differences in Enumerated Data Types. DEXA Workshops 2002: 883-888
- Cheng Hian Goh, Stéphane Bressan, Stuart E. Madnick, Michael Siegel: Context Interchange: New Features and Formalisms for the Intelligent Integration of Information. *ACM* Trans. Inf. Syst. 17(3): 270-293 (1999)
- Mark Hansen, Stuart E. Madnick, Michael Siegel: Data Integration using Web Services. DIWeb 2002: 3-16

Academic systems

Comparative Analysis

Evaluation criteria

- We evaluated the proposals with respect to 9 criteria in order to compare them by means of a set of synoptic headings
  1. Mapping representation (GAV / LAV)
  2. Types of sources considered
  3. Underlying Model / formalism / Logic
  4. Level of automation
  5. Technologies exploited for data integration (semantics level)
  6. Schema-based vs. Instance-based techniques
  7. Developed tool /Documented experiences of use
  8. Management of the source dynamics
  9. Future work

1. Mapping representation (GAV / LAV / GLAV)

- The approaches may represent mappings between the local sources and the global integrated schema by means of a Global as View or a Local as View approach

TSIMMIS	GARLIC	SIMS	Information Manifold	INFOMASTER	COIN	MOMIS
GAV	GAV	LAV	LAV	GAV	GAV	GAV

2. Types of sources considered

- The data integration systems may manage different kinds of data sources

TSIMMIS	GARLIC	SIMS	Information Manifold	INFOMASTER	COIN	MOMIS
structured, semi-structured	structured, semi-structured, multimedia	Data sources (Databases typically) represented by the LOOM logic	Structured, Semi-structured, OO sources	structured, semi-structured, WWW	Databases	Structured, Semi-structured, multimedia

3. Underlying Model / formalism / Logic

- The unified schema is represented by means an internal model. The sources model has to be translated into the unified model for the integration purposes
- Some logic techniques may be exploited for inferring new relationships between data sources and unified representation and for checking the consistency of the unified schema

TSIMMIS	GARLIC	SIMS	Information Manifold	INFOMASTER	COIN	MOMIS
OEM	GDL (ODMG 93 extension)	Object oriented and LOOM logic	Extended Relational and CARIN description Logics	KIF / KQML	Object oriented with axioms	Object oriented and OLCD descritpion logic

4. Level of automation

- The level of automation determines the system scalability and the ability of managing the evolutions of the sources

TSIMMIS	GARLIC	SIMS	Information Manifold	INFOMASTER	COIN	MOMIS
User supported	Extraction by means of automated tools, mappings built with the user interaction	Wrappers add sources in SIMS and the query process is based on description logic techniques	The use of automatic technique for extracting data is suggested	Not automatic	Not automatic: the user provides a logical specification of how data are interpreted and how conflicts have to be solved	Semi-automatic the adding of a new source, automatic the unified view building

5. Technologies exploited for data integration (semantics level)

- The techniques and the semantics exploited for extracting knowledge from the sources and for building the integrated schema are analyzed

TSIMMIS	GARLIC	SIMS	Information Manifold	INFOMASTER	COIN	MOMIS
Semantics Rules	Automatic RDBMS to ODMG wrappers	Loom DL and operations for selecting information sources, generalize concepts, specialize concepts, decompose-relation mainly based on schema analysis	Horn rules and Classic description logic; probabilistic formalism	No semantics, rules expressed by the user	Abduction inference	Lexical – structural Knowledge. Description Logic techniques

6. Schema-based vs. Instance-based techniques

- The techniques identified in the previous item may be applied (in principle) to the data schema or to the instances

TSIMMIS	GARLIC	SIMS	Information Manifold	INFOMASTER	COIN	MOMIS
schema based	schema based	schema based	schema based	schema based	schema based	schema based

7. Developed tool /Documented experiences of use

- We evaluate if there is for each approach a developed tool that provides some documented experiences of the system use in real domains

TSIMMIS	GARLIC	SIMS	Information Manifold	INFOMASTER	COIN	MOMIS
N.A.	the Garlic technology, particularly the wrapper and optimization technology, has been key to several IBM products and offerings	N.A.	N.A.	Some example of use in e-commerce. A commercial software has been obtained and commercialized	Papers describe the use in several business domain	Yes, in several scenarios

8. Management of the source dynamics

- Data sources evolve and a data integration system should manage such dynamics

TSIMMIS	GARLIC	SIMS	Information Manifold	INFOMASTER	COIN	MOMIS
It does not deal with this issue	It does not deal with this issue	No problem	No problem	It does not deal with this issue	It does not deal with this issue	Yes, techniques have been developed for checking when the unified schema has to be rebuilt

9. Future work

- All the analyzed system were developed more than 10 years ago. We analyzed if the project is still alive and/or which are the main outcomes

TSIMMIS	GARLIC	SIMS	Information Manifold	INFOMASTER	COIN	MOMIS
The project is ended on 1996	The project is ended on 1999 and continues in CLIO	Prometheus Framework for Data Integration (ISI Southern California university) and Ariadne (for web sources)	The project is ended on 1996	The project is ended on 1997	The project is ended on 1999	The project is still alive: it is investigating probabilistic techniques

Bibliography

- Yigal Arens and Chun-Nan Hsu and Craig A. Knoblock: Query Processing in the SIMS Information Mediator. 1997: 82--90
- Ilario Benetti, Domenico Beneventano, Sonia Bergamaschi, Francesco Guerra, Maurizio Vincini, "An Information Integration Framework for E-Commerce", IEEE Intelligent Systems Magazine, January/February 2002
- Domenico Beneventano, Sonia Bergamaschi, Francesco Guerra, Maurizio Vincini: "Synthesizing an Integrated Ontology ", IEEE Internet Computing Magazine, September-October 2003,42-51.
- Sonia Bergamaschi, Silvana Castano, Domenico Beneventano, Maurizio Vincini: "paper/Semantic Integration of Heterogeneous Information Sources"paper/, Special Issue on Intelligent Information Integration, Data & Knowledge Engineering, Vol. 36, Num. 1, Pages 215-249, Elsevier Science B.V. 2001.
- Sonia Bergamaschi, Silvana Castano, Maurizio 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.
- Philip A. Bernstein: The many roles of meta data in data integration. SIGMOD Conference 2005: 792
- Philip A. Bernstein, Sergey Melnik, Michalis Petropoulos, Christoph Quix: Industrial-Strength Schema Matching. SIGMOD Record 33(4): 38-43 (2004)
- Philip A. Bernstein, Laura M. Haas: Information integration in the enterprise. Commun. ACM 51(9): 72-79 (2008)
- Jens Bleiholder, Felix Naumann: Data fusion. ACM Comput. Surv. 41(1): (2008)

- Michael J. Carey and Laura M. Haas and Peter M. Schwarz and Manish Arya and William F. Cody and Ronald Fagin and John Thomas and John H and Edward L. Wimmers: Towards Heterogeneous Multimedia Information Systems: The Garlic Approach. 1995: 124—131
- Silvana Castano, Alfio Ferrara, Davide Lorusso, Stefano Montanelli: On the Ontology Instance Matching Problem. DEXA Workshops 2008: 180-184
- Silvana Castano, Valeria De Antonellis, Sabrina De Capitani di Vimercati: Global Viewing of Heterogeneous Data Sources. IEEE Trans. Knowl. Data Eng. 13(2): 277-297 (2001)
- Anish Das Sarma, Xin Dong, Alon Y. Halevy: Bootstrapping pay-as-you-go data integration systems. SIGMOD Conference 2008: 861-874
- AnHai Doan, Alon Y. Halevy: Semantic Integration Research in the Database Community: A Brief Survey. AI Magazine 26(1): 83-94 (2005)
- Luna Dong, Felix Naumann: Data fusion - Resolving data conflicts in integration, Tutorial at VLDB 2009
- Ron Fagin, Laura Haas, Mauricio Hernández, Renée J. Miller, Lucian Popa, Yannis Velegrakis: Clio: Schema Mapping Creation and Data Exchange To appear in book Conceptual Modeling: Foundations and Applications, editors Alexander Borgida, Vinay Chaudhri, Paolo Giorgini and Eric Yu, Springer 2009.
- Michael J. Franklin, Alon Y. Halevy, David Maier: From databases to dataspaces: a new abstraction for information management. SIGMOD Record 34(4): 27-33 (2005)
- Michael R. Genesereth and Arthur M. Keller and Oliver M. Duschka: Infomaster: An Information Integration System. 1997 ACM SIGMOD Conference: 539—542
- Fausto Giunchiglia, Mikalai Yatskevich, Pavel Shvaiko: Semantic Matching: Algorithms and Implementation. J. Data Semantics 9: 1-38 (2007)

- Cheng Hian Goh, Stéphane Bressan, Stuart E. Madnick, Michael Siegel: Context Interchange: New Features and Formalisms for the Intelligent Integration of Information. ACM Trans. Inf. Syst. 17(3): 270-293 (1999)
- Laura M. Haas: Beauty and the Beast: The Theory and Practice of Information Integration. ICDT 2007: 28-43
- Alon Y. Halevy: Data Integration: A Status Report. BTW 2003: 24-29
- Alon Y. Halevy: Learning about data integration challenges from day one. SIGMOD Record 32(3): 16-17 (2003)
- Alon Y. Halevy: Structures, Semantics and Statistics. VLDB 2004: 4-6
- Alon Y. Halevy, Naveen Ashish, Dina Bitton, Michael J. Carey, Denise Draper, Jeff Pollock, Arnon Rosenthal, Vishal Sikka: Enterprise information integration: successes, challenges and controversies. SIGMOD Conference 2005: 778-787
- Alon Y. Halevy, Anand Rajaraman, Joann J. Ordille: Data Integration: The Teenage Years. VLDB 2006: 9-16
- Alon Y. Halevy, Zachary G. Ives, Dan Suciu, Igor Tatarinov: Schema Mediation in Peer Data Management Systems. ICDE 2003: 505-516
- Hammer J., Stonebraker M., Topsakal O., THALIA: Test Harness for the Assessment of Legacy Information Integration Approaches. ICDE 2005: 485-486.
- Cheng Hian Goh and Stephane Bressan and Stuart Madnick and Michael Siegel: Context Interchange: New Features and Formalisms for the Intelligent Integration of Information. 1998

- Richard Hull: Managing Semantic Heterogeneity in Databases: A Theoretical Perspective. PODS 1997: 51-61
- W.H. Inmon: Building the data warehouse, Wiley Sons Publishing 1997
- Maurizio Lenzerini: Data Integration: A Theoretical Perspective. PODS 2002: 233-246
- Chen Li and Ramana Yerneni and Vasilis Vassalos and Hector Garcia-molina and Yannis Papakonstantinou and Jeffrey Ullman: Capability Based Mediation in TSIMMIS. ACM SIGMOD 1998: 564—566
- Jayant Madhavan, Philip A. Bernstein, AnHai Doan, Alon Y. Halevy: Corpus-based Schema Matching. ICDE 2005: 57-68
- Jayant Madhavan, Shirley Cohen, Xin Luna Dong, Alon Y. Halevy, Shawn R. Jeffery, David Ko, Cong Yu: Web-Scale Data Integration: You can afford to Pay as You Go. CIDR 2007: 342-350
- Sergey Melnik, Hector Garcia-Molina, Erhard Rahm: Similarity Flooding: A Versatile Graph Matching Algorithm and Its Application to Schema Matching. ICDE 2002: 117-128
- E. Mena and V. Kashyap and A. Sheth and A. Illarramendi: OBSERVER: An Approach for Query Processing in Global Information Systems based on Interoperation across Pre-existing Ontologies. IEEE Computer Society Press 1996: 14—25
- David Menestrina, Omar Benjelloun, Hector Garcia-Molina: Generic Entity Resolution with Data Confidences. CleanDB 2006
- Felix Naumann, Alexander Bilke, Jens Bleiholder, Melanie Weis: Data Fusion in Three Steps: Resolving Schema, Tuple, and Value Inconsistencies. IEEE Data Eng. Bull. 29(2): 21-31 (2006)
- Natalya Fridman : Semantic Integration: A Survey Of Ontology-Based Approaches. SIGMOD Record 33(4): 65-70 (2004)

•Yannis Papakonstantinou, Serge Abiteboul, Hector Garcia-Molina: Object Fusion in Mediator Systems. VLDB 1996: 413-424

•Rachel Pottinger, Philip A. Bernstein: Creating a Mediated Schema Based on Initial Correspondences. IEEE Data Eng. Bull. 25(3): 26-31 (2002)

•Rachel Pottinger, Philip A. Bernstein: Schema merging and mapping creation for relational sources. EDBT 2008: 73-84

•Erhard Rahm, Philip A. Bernstein: A survey of approaches to automatic schema matching. VLDB J. 10(4): 334-350 (2001)

•Pavel Shvaiko, Jérôme Euzenat: A Survey of Schema-Based Matching Approaches. J. Data Semantics IV: 146-171 (2005)

•Gio Wiederhold: Mediators in the Architecture of Future Information Systems. IEEE Computer 25(3): 38-49 (1992)

•Natalya Fridman Noy: Semantic Integration: A Survey Of Ontology-Based Approaches. SIGMOD Record 33(4): 65-70 (2004) 2003

• S. Bergamaschi, L. Po, S. Sorrentino: Automatic Annotation in Data Integration Systems, OTM Workshop 2007: 27-28

• S. Bergamaschi, L. Po, A. Sala, S. Sorrentino: Data source annotation in data integration systems, The Fifth International Workshop on Databases, Information Systems and Peer-to-Peer Computing (DBISP2P)

• S. Bergamaschi, L. Po, S. Sorrentino: Automatic annotation for mapping discovery in data integration systems, SEBD 2008: 334-341

•Laura Po: Improving Data Integration through Disambiguation Techniques. NLDB 2008: 372-375

•S. Bergamaschi, S. Sorrentino: Semi-automatic compound nouns annotation for data integration systems, SEBD 2009: 221-228

•L. Po, S. Sorrentino, S. Bergamaschi, D. Beneventano: Lexical Knowledge Extraction: an Effective Approach to Schema and Ontology Matching, ECKM 2009

• D. Beneventano, S. Bergamaschi, S. Sorrentino: Extending WordNet with compound nouns for semi-automatic annotation in data integration systems, IEEE NLP-KE 2009

•S. Bergamaschi, L. Po, S. Sorrentino, A. Corni: Uncertainty in data integration systems: automatic generation of probabilistic relationships, Itais 2009

•S. Sorrentino, S. Bergamaschi, M. Gawinecki, L. Po: Schema Normalization for Improving Schema Matching, to apper at ER 2009, 9-12 November, Gramado, Brasil.

•S. Bergamaschi, L. Po, S. Sorrentino, A. Corni: Dealing with Uncertainty in Lexical Annotation, to appear at ER 2009, 9-12 November 2009, Gramado, Brasil. 2009