Integration and Provenance of Cereals Genotypic and Phenotypic Data

Domenico Beneventano, Sonia Bergamaschi, Abdul Rahman Dannaoui

University of Modena and Reggio Emilia

20th Italian Symposium on Advanced Database Systems (SEBD)
June 24th - 27th 2012, Venice

Aim of the work

- The CEREALAB Database integrates data coming from different data sources by using the MOMIS Data Integration System, with the aim of creating a powerful tool for plant breeders and geneticists.
- A fundamental task in data integration is data fusion
- Provenance is one of the open problems and desiderata for data fusion systems. [Dong, Naumann, Data fusion: resolving data conflicts for integration. VLDB 2009]
- Provenance as a requirement emerging from CEREALAB users:
  - to know the origin of the visualized data
  - to explain merging decisions by tracking which original values were involeved and how they have been fused together.

➔ **Design and Development of a Provenance Managment component for the MOMIS System and its application to the CEREALAB domain.**

Table of content

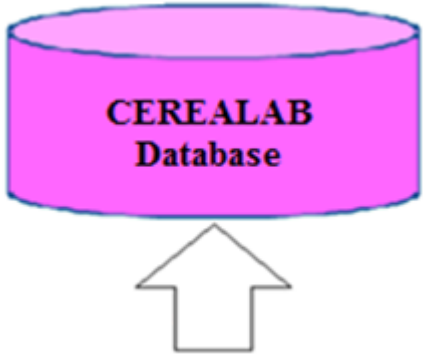
The MOMIS Data Integration System

• **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)

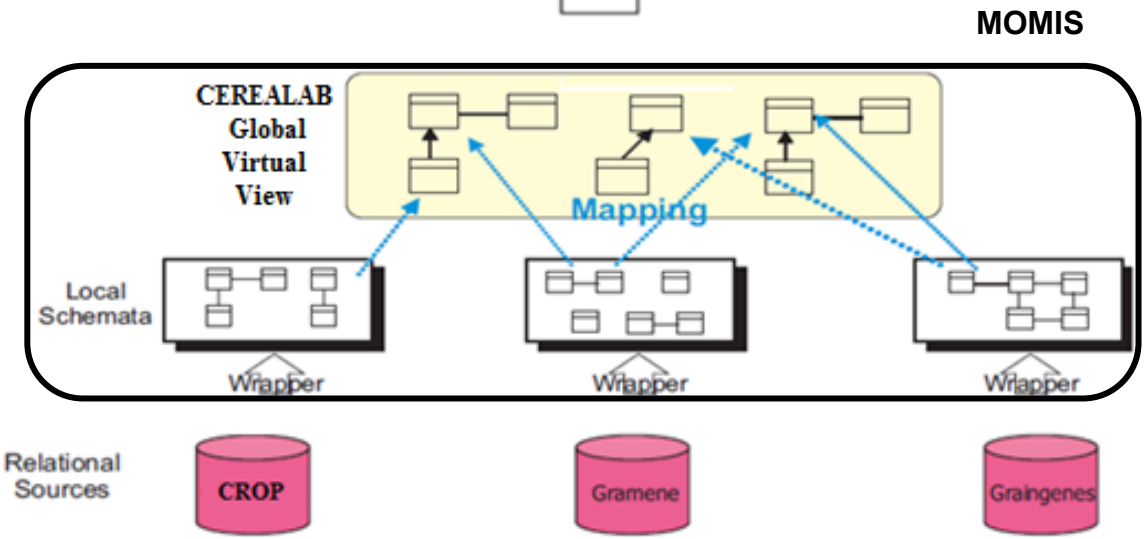
• **MOMIS OPEN SOURCE (MOMIS 1.2 Released this month)**
DataRiver (www.datariver.it) is a Start-Up company of the University of Modena and Reggio Emilia started on 17 June 2009 by professors and researchers of the DBGroup

Integration of biological data (CEREALAB DB)

The CEREALAB Database stores genotypic and phenotypic cereal data collected within the CEREALAB project and integrates them with already existing well known public data sources.

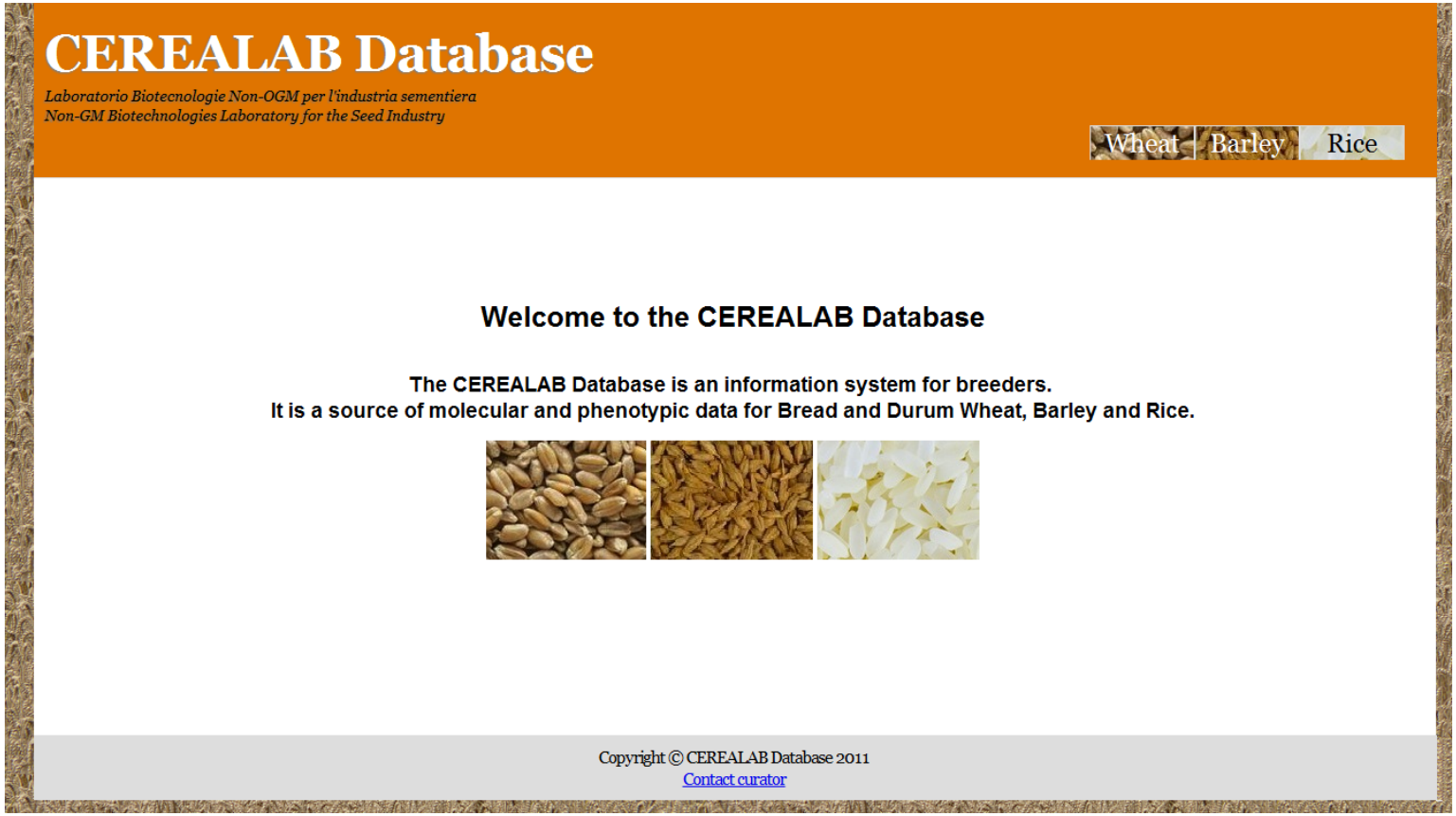


Data integration is obtained by using the MOMIS data integration system



Integration of biological data (CEREALAB DB)

http://www.cerealab.unimore.it/dbv3/



Example: **Germplasm** global class

Two local classes with two conflicting attributes

GPA (GermplasmA)

GPN	yield	FHB	Type
Eureka	18	MR	
Fortuna	7	MR	
Mentana		S	Line
Kenora	20	MR	Landrace
Oasis	21	MR	Cultivar

GPB (GermplasmB)

GPN	yield	FHB	Type
Eureka	6	S	Cultivar
Fortuna	15	S	Landrace
Mentana	20	MR	Line
Kenora			Cultivar

legend:

GPN : Shared identifier.
FHB : Fusarium Head Blight.
Yield: Production in t/ha.
Type: Germplasm type.

Data fusion

Full join merge operator (Resolution Functions for solving data conflicts):

```
SELECT    GPN=GPN,
          Yield=AVG(GPA.Yield,GPB.Yield),
          FHB= ALLVALUES(GPA.type,GPB.type)
FROM      GPA  FULL OUTER JOIN GPB
          USING (GPN)
```

GERMPLASM

GPN	yield	FHB	Type
Eureka	12	MR	Cultivar
Fortuna	11	MR	Landrace
Mentana	20	S	Line
Kenora	20	MR	Landrace, Cultivar
Oasis	21	MR	Cultivar

Provenace for the Full Join Merge Operator

- The "PI-CS provenance" gives more precise provenance information (w.r.t. other provenance models) for outerjoins

  [Glavic, Alonso. Perm: Processing provenance and data on the same data model through query rewriting, ICDE 2009]

- The provenance of the Full Join-Merge operator is defined by extending the concept of "*PI-CS* provenance" with Resolution Functions

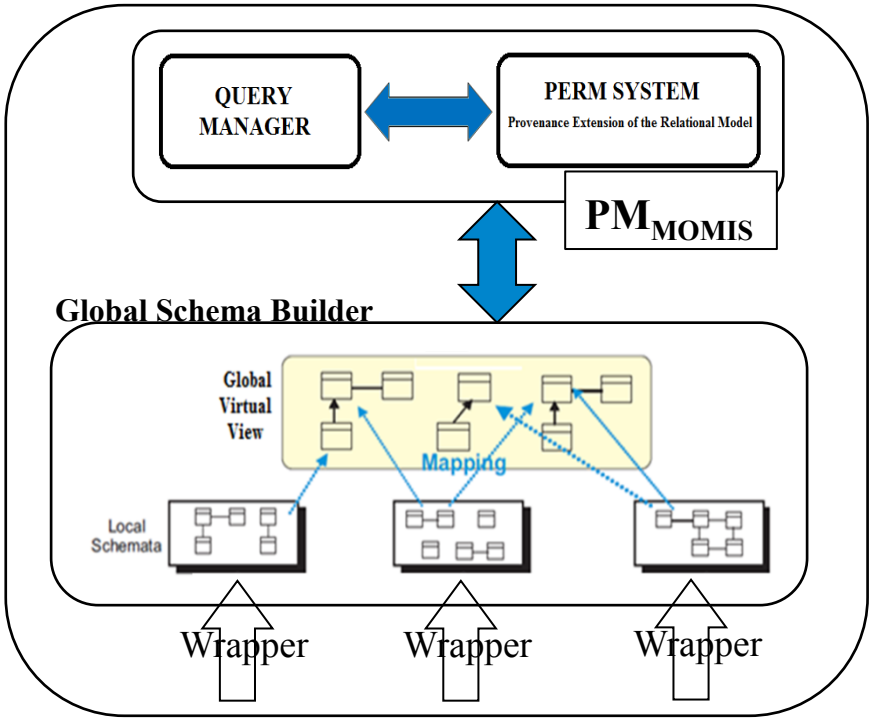
  [Beneventano et al. On Provenance of Data Fusion Queries. SEBD2011]

Provenace in the MOMIS System

- The "*PI-CS* provenance" is fully implemented in the "Perm" system, an open-source provenance management system.
  [Glavic, Alonso. The perm provenance management system in action, SIGMOD 2009]

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

- Extensions implemented to consider Resolution Functions (Provenance computation for the full join-merge)



Example: TYPE_MR query (1/3)

GERMPLASM Global Class

GPN	yield	FHB	Type
Eureka	12	MR	Cultivar
Fortuna	11	MR	Landrace
Mentana	20	S	Line
Kenora	20	MR	Landrace, Cultivar
Oasis	21	MR	Cultivar

Query: types of varieties that are resistant to FHB?

```
TYPE_MR= SELECT DISTINCT Type
         FROM   GERMPLASM
         WHERE  FHB='MR'
```

TYPE_MR query

Type
Landrance
Cultivar
Landrace,Cultivar

Example: TYPE_MR query (2/3)

➢ *P*rovenance for the TYPE_MR query

Type	*Provenance as a set of witness lists*
Landrance	{<GPA ^{Fortuna}, GPB ^{Fortuna}>}
Cultivar	{<GPA ^{Eureka}, GPB ^{Eureka}>, <GPA ^{Oasis}, ⊥>}
Landrace,Cultivar	{<GPA ^{Kenora}, GPB ^{Kenora}>}

➢ 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 (useful in modeling outerjoins).

Example: TYPE_MR query (3/3)

- 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

Type	A.GPN	A.yield	A.FHB	A.type	B.GPN	A.yield	B.FHB	B.type
landrace	Fortuna	7	MR		Fortuna	15	S	landrace
cultivar	Eureka	18	MR		Eureka	6	S	cultivar
cultivar	Oasis	21	MR	cultivar				
landrace,cultivar	Kenora	20	MR	landrace	Kenora			cultivar

Provenance based Conflict Handling Strategies

Query:
Germplasms of the same type of Kenora

GPN	...	Type
Eureka	...	Cultivar
Fortuna	...	Landrace
Kenora	...	Landrace, Cultivar

Problem:
The two values Landrace and Cultivar must be considered as *alternative values*; if not, we might obtain Fortuna and Eureka as germplasms of the same type.

➢The output of the full outer join merge operator is considered as an **uncertain relation** and is managed with a system that supports **uncertain data** and **data lineage**, the Trio system

[Benjelloun,O.,Sarma,A.D.,Hayworth,C.,Widom,J.:An introduction to ULDBs and the TRIO system. IEEE Data Eng. Bull. 29(1), 5–16 (2006)]

Conclusion and Future Work

✓ PM_{MOMIS} design and development. Some functionalities have been studied and partially implemented within the CEREALAB project.

• **Implementation**: to include some functionality of the Trio system (The Trio source code is freely available)

• **Other Provenance Models:**

1. Provenance semirings and the ORCHESTRA system.
   [Ives et al. Data integration and exchange for scientific collaboration. DILS 2009]

2. OPM (Open Provenance Model).
   [Moreau et al. The open provenance model: An overview.]

The MOMIS Data Integration System

- MOMIS detects semantic similarities among the involved local source schemata (*local classes*), groups semantic related local classes in global classes, thus obtaining a *Global Schema*

- **Mapping Table**: Correspondences among a global class and its local classes. Example: Global Class Hotel = { resort, hotel }

		resort	hotel
one-to-many	Name	name	name
	Room	rooms	hotelrooms
	Price	amount	price
one-to-one	Star	star	
	Wifi		wifi

- **Global-as-View** mappings: for each global class a query over its local classes is defined

- **Data Fusion**: the query is defined by a Full Join Merge operator

[Naumann et al. Completeness of integrated information sources. Inf. Syst. 2004]

Full Join Merge operator

Hypothesis: A shared identifier (ID) among all local classes.

1. **Full Join** on ID: all tuples of all local sources
2. **Merge** : data reconciliation (Resolution Functions for solving data conflicts in the case of one-to-many attributes)

Exampe for one-to-one mappings ➔ no conflicts

G	L1	L2
ID	ID	ID
A	A	
B		B

Intuitively, in SQL, G is defined as:

```
SELECT COALESCE(L1.ID,L2.ID) AS ID, L1.A AS A, L2.B AS B
FROM L1 FULL OUTER JOIN L2 ON (L1.ID=L2.ID)
```

L1

ID	A
1	3
2	3
3	ϵ
4	8

L2

ID	B
1	4
2	ϵ
3	ϵ
5	ϵ

G

ID	A	B
1	3	4
2	3	ϵ
3	ϵ	ϵ
4	8	ϵ
5	ϵ	ϵ

Full Join Merge: one-to-many mappings

Resolution Functions (RF) to solve conflicts for one-to-many global attributes

Intuitively, in SQL, G is defined as:

```
SELECT COALESCE(L1.ID,L2.ID) AS ID, L1.A  AS A, L2.B  AS B, RF (L1.C,L2.C) AS C
FROM L1 FULL OUTER JOIN L2 ON (L1.ID=L2.ID)
```

G	L1	L2
ID	ID	ID
A	A	
B		B
C	C	C

L1

ID	A	C
1	3	4
2	3	3
3	ϵ	3
4	8	3

L2

ID	B	C
1	4	2
2	ϵ	ϵ
3	ϵ	3
5	ϵ	ϵ

G

ID	A	B	C=AVG(L1.C,L2.C)
1	3	4	3
2	3	ϵ	3
3	ϵ	ϵ	3
4	8	ϵ	3
5	ϵ	ϵ	ϵ