DBGroup@unimo

Virtual Integration of Existing Web Databases: Genotypic Selection of Cereal Cultivars

Sonia Bergamaschi, Antonio Sala
Dipartimento di Ingegneria dell'Informazione
Università di Modena e Reggio Emilia
bergamaschi.sonia@unimore.it, sala.antonio@unimore.it

Motivations and Domain

To perform intelligent data integration of existing databases to create a Global Virtual View(GVV) for the genotypic selection of cereal cultivars.

The GVV has been realized with the **MOMIS** system (Mediator envirOnment for Multiple Information Sources) (http://dbgroup.unimo.it/Momis/) developed by the Database Group of the University of Modena and Reggio Emilia as a part of the CEREALAB project conducted by the Agrarian faculty of the University of Modena and Reggio Emilia in collaboration and funded by the Regional Government of Emilia Romagna.

Integration Process Overview

The process gives rise to Global Virtual View of several specific data sources.
The steps of the Integration Process were:
- Insertion of a pre-existing local source
- Local source schemata extraction (Gramene and Graingenes databases)
- Local source annotation with WordNet
- Common Thesaurus generation
- GVV generation
- Mapping refinement

The ODL_{i3} language

MOMIS uses an object-oriented language called ODL_{i3} as a common data model for integrating a given set of local information sources.
ODL_{i3} extends ODL with the following relationships expressing intra- and inter-schema knowledge for the source schemata:
SYN (synonym of),
BT (broader terms),
NT (narrower terms),
RT (related terms).
By means of ODL_{i3}, only one language is exploited to describe both the sources (the input of the synthesis process) and the GVV (the result of the process).
ODL_{i3} is based on the OCDL description logics. Translators ODL_{i3}/OCDL and OCDL/ODL_{i3} are available.

Local source schemata extraction

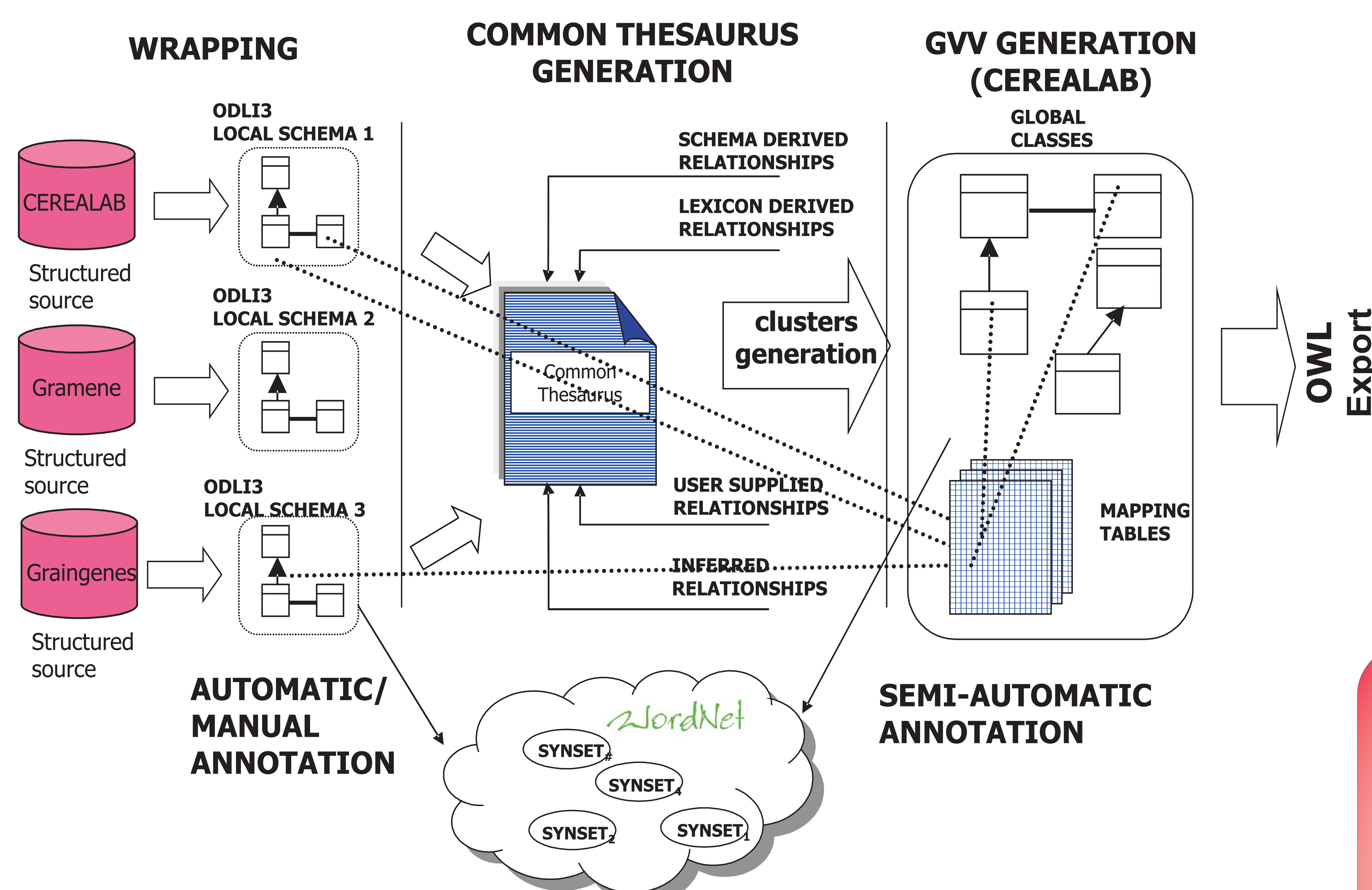
Choice of the data sources and their translation into ODL_{i3} format.
A pre-defined ontology existed. It has been enriched by other data sources, Gramene (http://www.gramene.org) and Graingenes (http://wheat.pw.usda.gov), chosen as they are considered the most significant for the domain.

The MOMIS wrappers logically converts the source schema description into an equivalent ODL_{i3} schema.
After this step we have three local sources: the pre-existing ontology (CEREALAB), and the Gramene and Graingenes sources

Local Source Annotation

- Assign a name and a set of meanings belonging to the WordNet lexical system to each local class and attribute of the local schemata.
  For each element of a local schema the system automatically suggests a word form corresponding to the given term (if it exists): the designer may confirm or change the word form or meaning of each element.
- MOMIS provides the user with a WordNet Editor to extend WordNet by adding new terms and synsets to the native elements of WordNet.

This extension step has to be performed just the first time a domain is handled.



Common Thesaurus Generation

MOMIS constructs a Common Thesaurus (CT) describing intra and inter-schema knowledge in the form of SYN (synonyms), BT/NT(broader terms/narrower terms), and RT (meronymy/holonymy) relationships among local schema elements.
The Common Thesaurus is constructed through an incremental process in which the following relationships are added:
- schema-derived relationships: relationships holding at intra-schema level are automatically extracted by analyzing each schema separately. For example, MOMIS extracts intraschema RT relationships from foreign keys in relational source schemas. When a foreign key is also a primary key, in both the original and referenced relation, MOMIS extracts BT and NT relationships, which are derived from inheritance relationships in object-oriented schemas.
- lexicon-derived relationship: we exploit the annotation phase in order to translate relationships holding at the lexical level into relationships to be added to the Common Thesaurus.
- designer-supplied relationships: new relationships can be supplied directly by the designer, to capture specific domain knowledge.
- inferred relationships: Description Logics (DL) techniques of ODB-Tools (http:/www.dbgroup.unimo.i/tODB-Tools.html) are exploited to infer new relationships, by means of subsumption computation applied to a ``virtual schema'' obtained by interpreting BT/NT as subclass relationships and RT as domain attributes.

Global Virtual View Generation

MOMIS
- identifies and groups similar ODL_{i3} classes (classes that describe the same or semantically related concept in different sources) into clusters (global classes)
- Generates mappings among global and local classes in the cluster

  Cluster generation: affinity coefficients are evaluated for all possible pairs of ODL_{i3} classes, based on the relationships in the Common Thesaurus properly strengthened
  Affinity coefficients determine the degree of matching of two classes based on:
  - their names (Name Affinity coefficient)
  - their attributes (Structural Affinity coefficient)

  Affinity coefficients are fused into Global Affinity coefficients calculated by means of the linear combination of the two coefficients.
  Global affinity coefficients are used by a hierarchical clustering algorithm, to include ODL_{i3} classes in clusters according to their degree of affinity.

The designer may interactively refine and complete the proposed integration results
  the mappings which has been automatically created by the system can be fine tuned.

Mapping Refinement

A Mapping Table (MT) is automatically generated for each global class of a GVV.
The designer can extend the MT by adding:
- Data Conversion Functions from local to global attributes
  The Ontology Designer can define, for each not null element, a Data Conversion Function which represents the mapping of local attributes into the global attribute
- Join Conditions among pairs of local classes.
  To identify instances of the same object and fuse them we introduce Join Conditions among pairs of local classes belonging to the same global class.
- Resolution Functions for global attributes to solve data conflicts of local attribute values.
  MOMIS provides some standard kinds of resolution functions for solving data conflicts for each global attribute mapping onto local attributes coming from more than one local source:
  - Random
  - Aggregation
  - Coalescence
  - Precedence function
  - All Values

DBGroup@unimo

Virtual Integration of Existing Web Databases: Genotypic Selection of Cereal Cultivars

Sonia Bergamaschi, Antonio Sala
Dipartimento di Ingegneria dell’Informazione
Università di Modena e Reggio Emilia
bergamaschi.sonia@unimore.it, sala.antonio@unimore.it

Querying the Global Virtual View with the MOMIS Query Manager

The **MOMIS** Query Manager is the coordinated set of functions which

- takes an incoming query (say global query),
- defines a decomposition of the query according to the mapping of the GVV onto the local data sources
- sends the subqueries to these data sources
- collects their answers
- fuse them (performing any residual filtering as necessary)
- delivers the answer

Query processing consists of the following steps:

- Query rewriting
  to rewrite a global query as an equivalent set of queries expressed on the local sources (local queries)
- Local queries execution
  the local queries are sent and executed at local sources
- Fusion and Reconciliation
  The local answers are fused into the global answer

An example **Query**

```
SELECT t.trait_name, q.name,
q.chromosomearm, m.marker_name
FROM trait_affected_by_qtl as t, qtl as q,
marker_for_qtl as m
WHERE t.trait_name LIKE '%powdery mildew%'
and t.qtl_name=q.name
and q.name=m.qtl_name
```

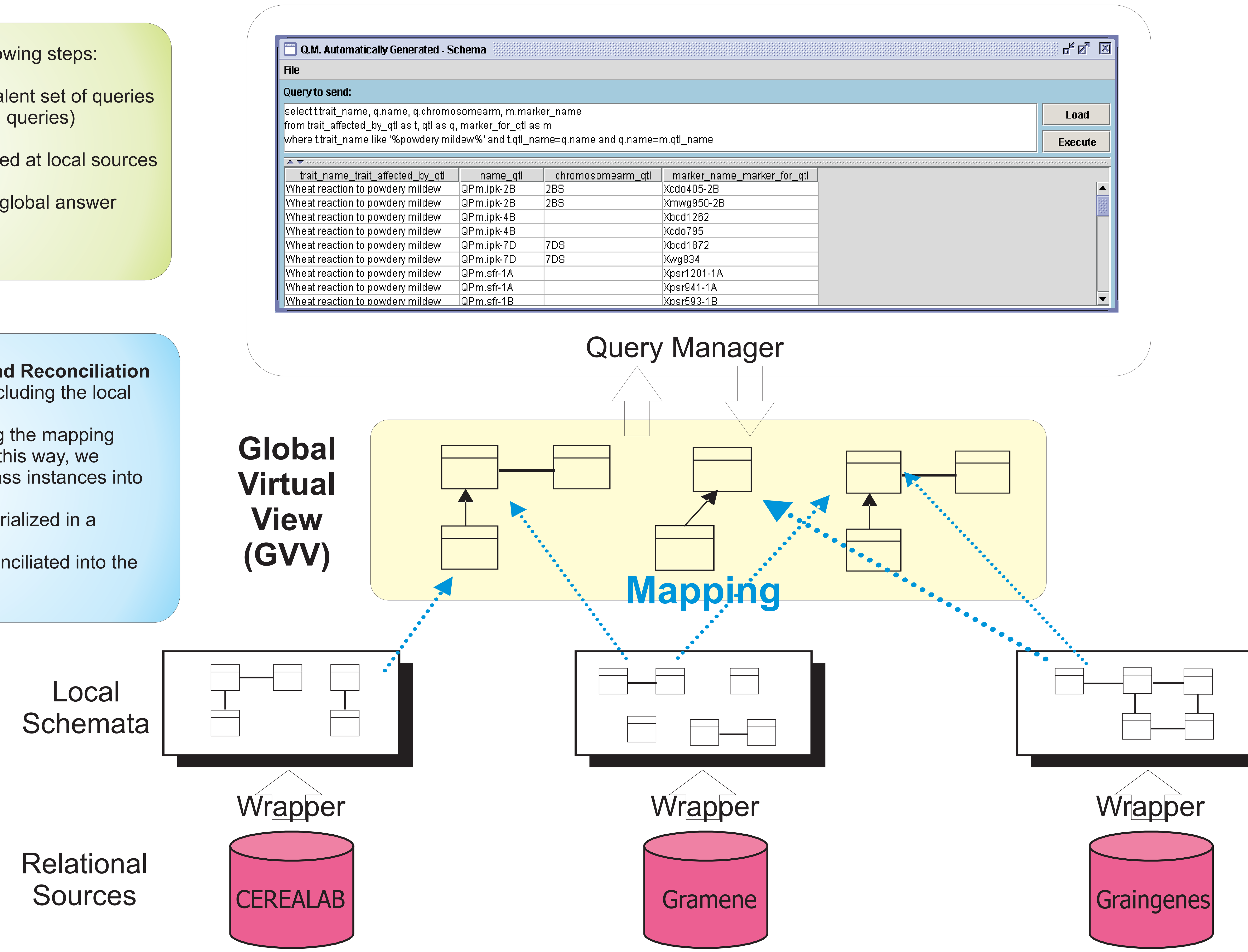
Is rewritten by means of unfolding (expanding each atom of the global query according to its definition in the mapping)

Local Queries Execution, Fusion and Reconciliation

- A local query is sent to the source including the local class
- Its answer is transformed by applying the mapping functions related to the local class: in this way, we perform the conversion of the local class instances into the GVV instances.
- The result of this conversion is materialized in a temporary table.
- Temporary tables are fused and reconciliated into the global answer.

Query rewriting process

- Atomic constraint mapping
  each atomic constraint of a query is rewritten into one that can be supported by the local class. The atomic constraint mapping is performed on the basis of mapping functions defined in the Mapping Table
- Residual Constraints computation
  residual constraints are the constraints of the global query that are not mapped in all local queries
- Local select-list computation
  The select-list of a local query is a set of attributes, including the global query attributes, the join attributes, the residual constrains attributes, translated into the correspondent set of local attributes on the basis of the mapping table.



trait_name_trait_affected_by_qtl	name_qtl	chromosomearm_qtl	marker_name_marker_for_qtl
Wheat reaction to powdery mildew	QPm.ipk-2B	2BS	Xcdo405-2B
Wheat reaction to powdery mildew	QPm.ipk-2B	2BS	Xmwg950-2B
Wheat reaction to powdery mildew	QPm.ipk-4B		Xbcd1262
Wheat reaction to powdery mildew	QPm.ipk-4B		Xcdo795
Wheat reaction to powdery mildew	QPm.ipk-7D	7DS	Xbcd1872
Wheat reaction to powdery mildew	QPm.ipk-7D	7DS	Xwg834
Wheat reaction to powdery mildew	QPm.sfr-1A		Xpsr1201-1A
Wheat reaction to powdery mildew	QPm.sfr-1A		Xpsr941-1A
Wheat reaction to powdery mildew	QPm.sfr-1B		Xpsr593-1B