DBGroup@unimo

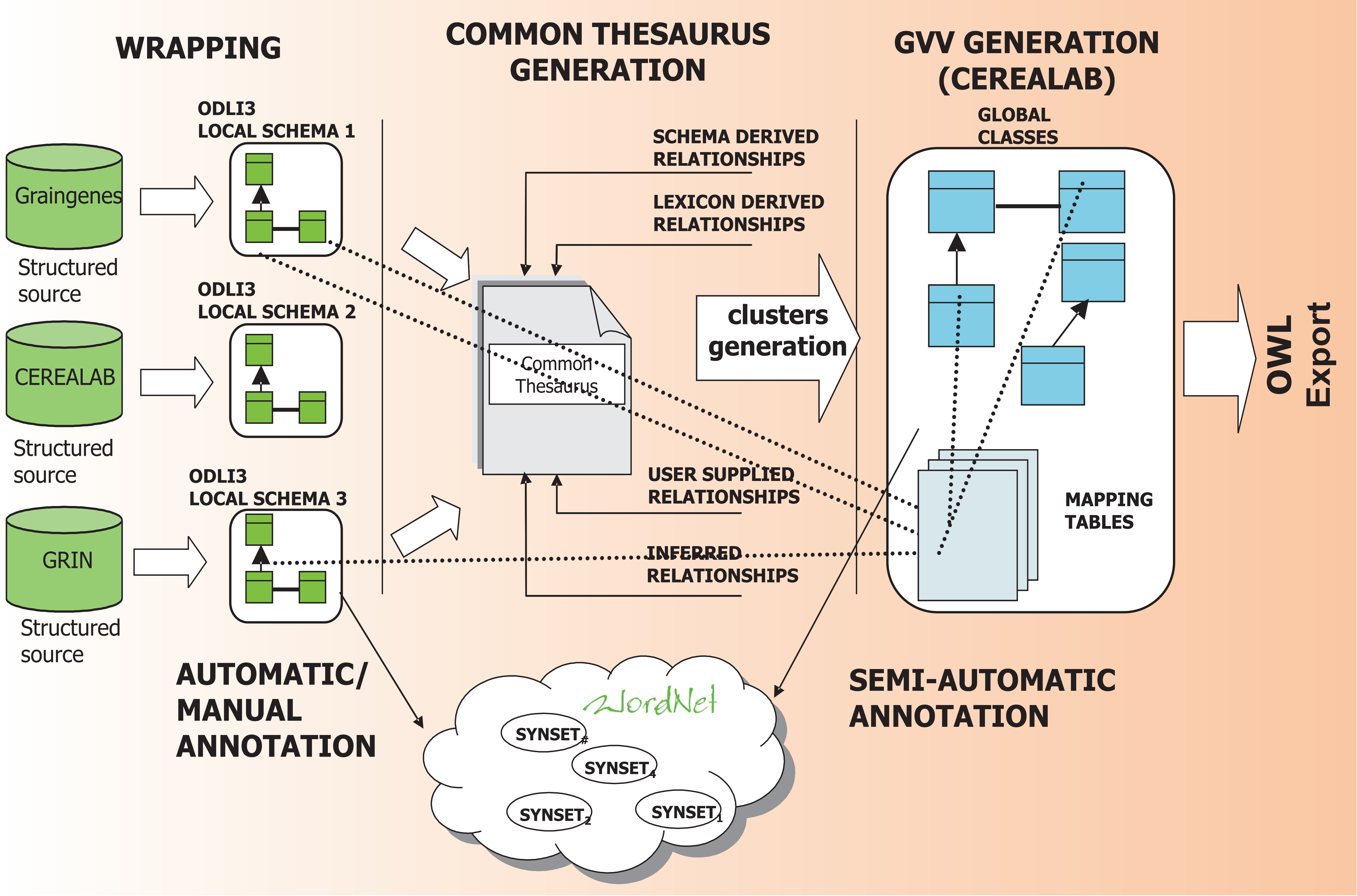
CEREALAB Database: Data Integration with the MOMIS System

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

Abstract

Biological informations are frequently widespread over the Web and retrieving knowledge in this domain often requires to navigate through several websites. Data sources are usually heterogeneuos and present different structures and interfaces. Mediator systems can be used to perform integration of such databases in order to have integrated view of multiple information sources and to query them.
The MOMIS system (Mediator envirOnment for Multiple Information Sources) is a framework developed by the Database Group of the University of Modena and Reggio Emilia (www.dbgroup.unimo.it) to perform intelligent information integration from both structured and semistructured data sources.
The result of the integration process is a Global Virtual View (GVV) of the underlying sources which is a conceptualization of the underlying domain and then may be thought of as an ontology describing the involved sources.
Queries can be posed over the GVV regardless of the structure of the local sources in a transparent way for the user.

The MOMIS System has been experimented for the realization of the CEREALAB database.
CEREALAB is a research project of technology transfer for applying Marker Assisted Selection (MAS) techniques to Cereal Breeding in Italian Seed Companies.



The ODL$_{i3}$ language

An object-oriented language called ODL$_{i3}$ is used as a common data model for integrating a given set of local information sources.

ODL$_{i3}$ extends ODL with the relationships expressing intra- and inter-schema knowledge for the source schemata:
SYN (synonym of),
BT (broader terms),
NT (narrower terms),
RT (related terms).

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 OLCD description logics.

Local source schemata extraction and Annotation

The MOMIS wrappers logically converts the source schema description into an equivalent ODL$_{i3}$ schema
In the Annotation phase we assign a name and a set of meanings belonging to the WordNet lexical system to each local class and attribute of the local schemata.

A word form corresponding to the given term is suggested by the system (if it exists): the designer may confirm or change the word form or meaning of each element.

MOMIS provides 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

intra and inter-schema knowledge is described in a Common Thesaurus built by MOMIS
SYN (synonyms)
BT/NT(broader terms/narrower terms)
RT (meronymy/holonymy)
relationships among local schema elements.

The Common Thesaurus is constructed adding incrementally:
✎schema-derived relationships:
relationships at intra-schema level extracted by analyzing each schema separately.
e.g.: 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, BT and NT relationships are derived from inheritance relationships in object-oriented schemas.
✎lexicon-derived relationship:
the annotation phase is exploited to translate relationships holding at the lexical level
✎designer-supplied relationships:
new relationships can be supplied directly by the designer
✎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 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.
Some standard kinds of resolution functions are provided for solving data conflicts for each global attribute coming from more than one local source.

Querying the Global Virtual View with the MOMIS Query Manager

The result of the integration process is a global virtual schema (GVV) of the underlying sources which can be thought as an ontology regarding both phenotypic and genotypic information about cereals.

A graphical interface has been developed for supporting visual query formulation: visual query are automatically translated into SQL queries supported by the MOMIS Query Manager. This interface is useful for users, like biologists, who do not have specific information technology skills.
In this way users can have access to data coming from different data source through a single interface without taking care about the structure or the query interface of each single database.
The GVV can be exported in OWL thus guaranteeing interoperability with other external applications/ontologies or external users

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:

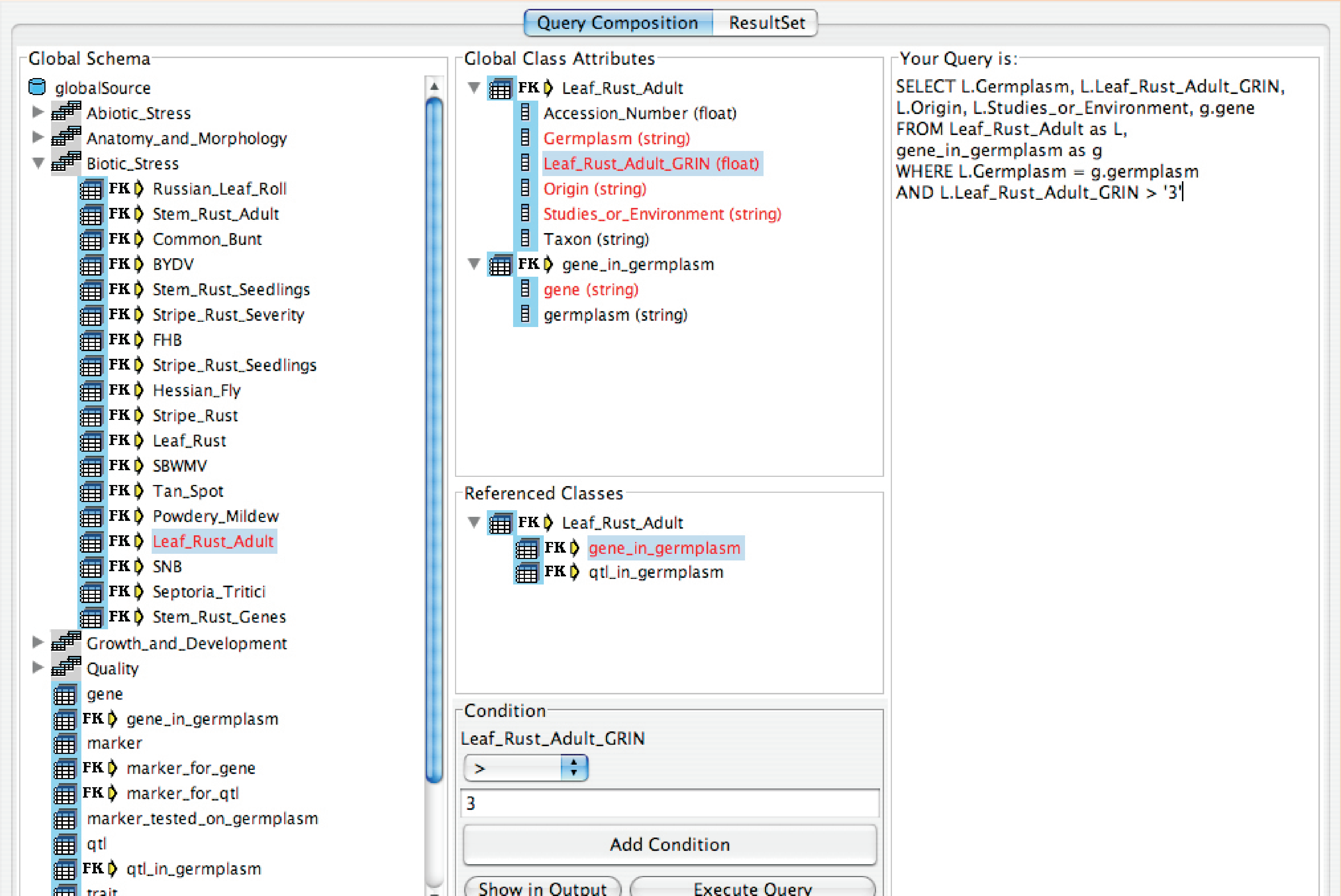
- Query rewriting
  to rewrite a global query as an equivalent set of queries expressed on the local sources (local queries)
- Local queries execution
  the local queries are sent and executed at local sources
- Fusion and Reconciliation
  The local answers are fused into the global answer

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.

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.



Global Schema

globalSource
- Abiotic_Stress
- Anatomy_and_Morphology
- Biotic_Stress
  - FK Russian_Leaf_Roll
  - FK Stem_Rust_Adult
  - FK Common_Bunt
  - FK BYDV
  - FK Stem_Rust_Seedlings
  - FK Stripe_Rust_Severity
  - FK FHB
  - FK Stripe_Rust_Seedlings
  - FK Hessian_Fly
  - FK Stripe_Rust
  - FK Leaf_Rust
  - FK SBWMV
  - FK Tan_Spot
  - FK Powdery_Mildew
  - FK Leaf_Rust_Adult
  - FK SNB
  - FK Septoria_Tritici
  - FK Stem_Rust_Genes
- Growth_and_Development
- Quality
- gene
- FK gene_in_germplasm
- marker
- FK marker_for_gene
- FK marker_for_qtl
- marker_tested_on_germplasm
- qtl
- FK qtl_in_germplasm
- trait

Global Class Attributes

- FK Leaf_Rust_Adult
  - Accession_Number (float)
  - Germplasm (string)
  - Leaf_Rust_Adult_GRIN (float)
  - Origin (string)
  - Studies_or_Environment (string)
  - Taxon (string)
- FK gene_in_germplasm
  - gene (string)
  - germplasm (string)

Referenced Classes

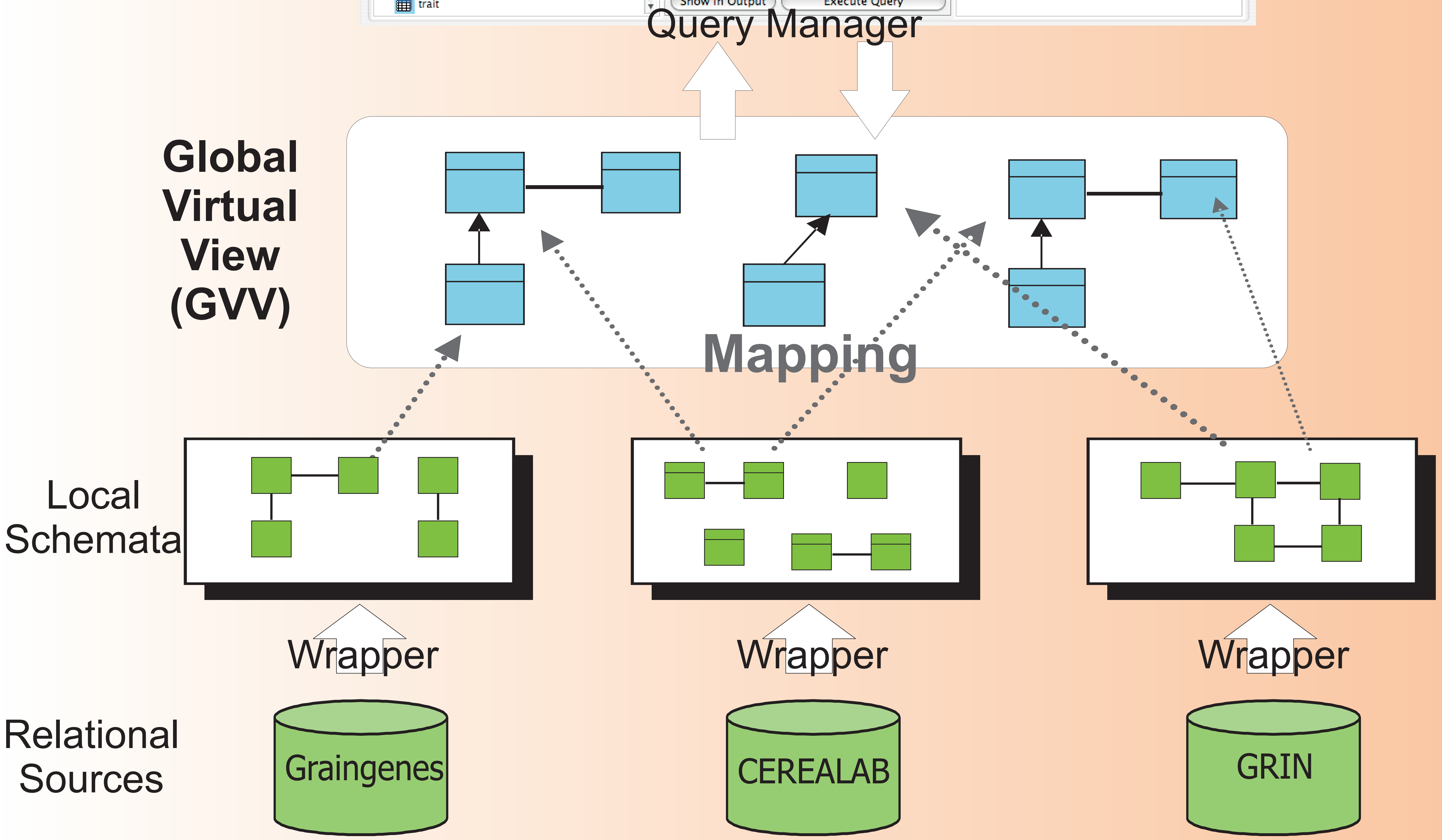
- FK Leaf_Rust_Adult
  - FK gene_in_germplasm
  - FK qtl_in_germplasm

Condition

Leaf_Rust_Adult_GRIN: > 3

Your Query is:

```
SELECT L.Germplasm, L.Leaf_Rust_Adult_GRIN,
L.Origin, L.Studies_or_Environment, g.gene
FROM Leaf_Rust_Adult as L,
gene_in_germplasm as g
WHERE L.Germplasm = g.germplasm
AND L.Leaf_Rust_Adult_GRIN > '3'
```


Query Manager
Global Virtual View (GVV)
Mapping
Local Schemata
Wrapper
Wrapper
Wrapper
Relational Sources
Graingenes
CEREALAB
GRIN