A Mediator Based Approach to Ontology Generation and Querying of Molecular and Phenotypic Cereals Data

Antonio Sala*
DBGroup - Dipartimento di Ingegneria dell'Informazione
Via Vignolese 905
Modena, Italy
antonio.sala@unimore.it
*Corresponding author

Sonia Bergamaschi
DBGroup - Dipartimento di Ingegneria dell'Informazione
Via Vignolese 905
Modena, Italy
sonia.bergamaschi@unimore.it

Abstract: In this paper we describe the development of the CEREALAB ontology, an ontology of molecular and phenotypic cereals data, realized by integrating public web databases with the database developed by the research group of the CEREALAB laboratory. Integration is obtained by using the MOMIS system (Mediator envirOnment for Multiple Information Sources), a data integration system developed by the Database Group of the University of Modena and Reggio Emilia. Information integration is performed in a semi-automatic way creating a Global virtual Schema (GS), that can be seen as an ontology of the underlying data sources, for which mapping rules and integrity constraints are specified to handle heterogeneity. The GS can be queried transparently w.r.t. the integrated data sources by using an easy-to-use graphical interface regardless of the specific languages of the source databases, and allows combining molecular data with the phenotypic information of cereals, to identify the correlation between the phenotype of a plant with the molecular data that express it.

Keywords: CEREALAB; ontology; molecular data; phenotypic data; cereals; data integration; mediator; querying.

Biographical notes: Antonio Sala is a Ph.D. Student at the International Doctorate School in Information and Communication Technologies of the University of Modena and Reggio Emilia. He's member of the DBGroup, the database research group of the same University (www.dbgroup.unimo.it). His research interests are mainly devoted to intelligent information integration related in particular to its application to biological databases.

Sonia Bergamaschi is Full Professor of Databases at the University of Modena and Reggio Emilia and Dean of the International Doctorate School in Information and Communication Technologies of the University of Modena and Reggio Emilia. She leads the database research group, DBGroup. Her research activities include data management, knowledge representation, reasoning techniques applied to databases, Semantic Web and data integration systems.

1 INTRODUCTION AND MOTIVATION

In the last few years numerous public data sources have been realized and made available for researchers in the field of molecular biology.

The main problem is that these data sources have different and heterogeneous structures and interfaces, and a different way of presenting their data. Moreover, the users are typically biology researchers with low information technology skills. As a consequence, a simple information search may take long time and eventually fail, even because of the number of different data sources to be accessed. What users need is, thus, to have access to the information available in different data sources in a transparent and easy

way, independently from the format or languages of the different sources.

Molecular biology data can be divided in genotypic data, concerning an organism's full hereditary information, even if not expressed; and phenotypic data, concerning an organism's actual observed properties, such as morphology, development, or behaviour. In particular, molecular data for cereals are available in different public reference databases: for example, Graingenes is considered one of the most important data source about wheat and barley, while Gramene can be considered one of the most complete data source about rice.

Concerning phenotypic data about cereals, i.e. the observable characteristic of the plants, the above-mentioned databases present also descriptions of phenotypic characters, but no quantitative evaluation of such traits is available. These kinds of data are available in specific data sources, such as the American Germplasm Resources Information Network (GRIN), which provides phenotypic information about many germplasms. Another valuable data source for phenotypic data is the Italian Council of Research in Agriculture (CRA), which collects and makes publicly available hundreds of observations of cereals traits in Italy.

It would be desirable to combine the molecular data with the phenotypic information of the plants, to identify the correlation between the phenotype of a plant with the molecular data that express it.

The aim of our work is thus to create a unique ontology (i.e. a Global virtual Schema, GS) of both molecular and phenotypic data about wheat, barley and rice, with an easy-to-use, transparent interface to the above-mentioned public data sources. This work was part of the activities of the CEREALAB laboratory, conducted jointly by the Agrarian faculty and the Database Group of the University of Modena and Reggio Emilia funded by the Regional Government of Emilia Romagna. The aim of the CEREALAB laboratory is to make available to the cereal breeders of the Emilia Romagna region a tool to perform genotypic selection of cereal cultivars from phenotypic traits. Since CEREALAB laboratory performs also genotyping activities, the integrated ontology has to allow the integration of new molecular data coming from the research activity carried out in the laboratory. The ontology is obtained by means of a mediator-based system (Wiederhold, 1992), called MOMIS and developed by the DBGroup of the University of Modena and Reggio Emilia. MOMIS allows integrating different and heterogeneous data sources creating a Global Virtual Schema (GS) that can be seen as a domain ontology that emerges from the schemas of the sources being integrated. This paper is more focused on the result of the integration process rather than the integration process itself. The integration process will be briefly described in the following, for a complete description see Bergamaschi, Castano, and Vincini (1999), Bergamaschi et al. (2001), Bergamaschi and Sala (2006).

As far as we know, no resource is available containing both molecular and phenotypic data about cereals suitable for the purpose of the CEREALAB laboratory. For this reason, we developed a GS that is the integration of existing molecular and phenotypic data sources with data provided by the CEREALAB laboratory. The GS can thus be seen as an ontology of the underlying sources regarding wheat, barley and rice.

Moreover, an important requirement we addressed in our work is usability: as this ontology is a working tool for users with high domain knowledge and low information technology expertise, it follows that the usage of the system has to be as much user-friendly as possible. For this reason we developed a graphical interface that can be used to query the ontology just pointing and clicking the concepts we are interested in, with no need to know any specific language usually needed for query formulation.

In this paper we present the CEREALAB ontology and its interface, and sketch the translation of graphical queries into queries executable by the Query Manager of the MOMIS system.

The rest of the paper is organized as follows: Section 2 presents some related work, Section 3 describes the domain of the CEREALAB laboratory to clarify the terms used and the data sources involved in the integration process. Section 4 briefly presents the MOMIS system and the data integration process. Section 5 describes the CEREALAB ontology obtained while Section 6 sketches out the querying process with the MOMIS Query Manager and presents the user interface developed to graphically formulate SQL queries over the ontology. Finally, Section 7 gives conclusions and future works.

2 RELATED WORK

The research community has produced a large number of projects, systems and techniques aiming at integrating heterogeneous data sources. A common approach for integrating information sources is to build a mediated schema as a synthesis of them (Wiederhold, 1992). By managing all the collected data in a common way, a mediated schema allows the user to pose a query according to a global perception of the handled information. A query over the mediated schema is translated into a set of sub-queries for the involved sources by means of automatic unfolding-rewriting operations taking into account the mediated and the sources schemata. Results from sub-queries are finally unified by data reconciliation techniques. A survey of several systems that follow this approach is available in Halevy et al. (2006).

The problem of data integration for biology has become really important in the last few years both due to continuous increases in data volumes and the growing diversity in types of data to be managed. Ontologies have been identified as a possible solution to the data integration needs in this field. The Open Biomedical Ontologies (OBO) Foundry is a collaborative experiment involving developers of science-based ontologies who are establishing a set of principles for ontology development with the goal of creating a suite of

orthogonal interoperable reference ontologies in the biomedical domain. In the OBO Foundry, some ontologies related to the domain of the CEREALAB laboratory exist, but none of these correlates phenotypic data with molecular data. The most important one is the Cereal Plant Trait Ontology (TO), which is a controlled vocabulary that describes each trait as a distinguishable feature, characteristic, quality or phenotypic feature of a developing or mature individual. The TO partially covers our domain of interest, and thus has been used as a reference, but the ontology we developed overcomes the TO as it integrates the trait ontology with molecular data related to phenotypic data.

Concerning the problem of data integration for biology, many research efforts have been devoted to developing systems that allow integrating different data sources. For example the Transparent Access to Multiple Bioinformatics Information Sources project, known as TAMBIS (Stevens et al., 2000), is a mediator-based integration system in which a domain ontology for molecular biology and bioinformatics is used in a retrieval-based information integration system for biologist. TAMBIS uses the global ontology to formulate queries through a graphical interface where a user needs to browse through concepts defined in a global schema and select the ones that are of interest for the particular query.

The Semantic Meta Database, SEMEDA (Köhler et al., 2003), is an ontology based integration framework for molecular biological data sources that uses its own ontology for database integration. This ontology is a small top-level ontology, which defines databases at the schema level. The idea behind the top-level structure of SEMEDA's ontology is that attributes in molecular biological databases generally store names, identifiers, properties and free text descriptions of real world objects. Names and identifiers serve to identify real world objects, whereas properties and descriptions store facts about those objects. Out of this, 'name', 'identifier', 'description' and 'property' are used as top-level concepts in SEMEDA's ontology. Database tables and attributes are then manually mapped to the given ontology. A query interface guides the user through the ontology to the relevant database/tables attributes and support the construction of the query.

In both the TAMBIS and the SEMEDA approaches, the global ontology is designed manually (i.e. the integration designer has to manually add the concepts to the global ontology appearing in the data sources that need to be integrated) while with the MOMIS system the ontology development is a semi-automatic process, i.e. the ontology emerges from the integration process as the Global Schema of the underlying source schemas. Another difference is in the fact that in TAMBIS and SEMEDA, mappings among the Global Schema and the local sources are defined manually, while in the MOMIS system mappings and clusters of similar classes are automatically generated once the sources have been semi-automatically, lexically annotated, as shown in section 4. The generation of the GS is thus semi-automatic.

BioKleisli (Davidson et al., 1997) is primarily a loosely-coupled federated database system. The mediator on top of the underlying integration system relies mainly on a high-level query language (the Collection Programming Language, or CPL), more expressive than SQL, that provides the ability to query across several sources. The BioKleisli project is mainly aimed at performing a horizontal integration. In fact, a query attribute is usually bound to an attribute in a single predetermined source; there is essentially no integration of sources with content overlap. K2 (Davidson et al., 2001) is the newer version of the BioKleisli system. K2 abandons CPL and replaces it by OQL, a more widely used query language. This change does not modify the overall flow of the system. Queries are still decomposed into subqueries and sent to the underlying sources using data drivers, while the query optimizer remains a rule-based optimizer. DiscoveryLink (Haas et al., 2001) is a mediator-based and wrapper-oriented middleware integration system. It serves as an intermediary for applications that need to access data from several biological sources. Applications typically connect to DiscoveryLink and submit a query in SQL on the global schema, not necessarily aware of the underlying sources.

These two systems offer format and location transparency but do not offer data reconciliation.

A survey of these and some other well-known systems that are currently available can be found in (Hernandez and Kambhampati, 2004).

As it can be seen, the data integration problem for biology has been addressed in numerous ways, but as far as we know the approach presented in this paper is the first one that has the capability to integrate overlapping data sources and to semi-automatically build the Global Schema. Moreover, the CEREALAB ontology developed is a new result in terms of its content as it combines molecular and phenotypic data. In fact, all the mentioned systems integrate only molecular data sources. Finally, except TAMBIS, SEMEDA and MOMIS, the existing systems use some kind of query languages instead of a graphical user interface to formulate queries. Easy query formulation is a pre-requisite for users of this kind of systems who usually have low IT expertise and thus need a user-friendly system.

3 DESCRIPTION OF THE DOMAIN AND OF THE DATA SOURCES

3.1 The CEREALAB Domain

The reference domain for our work regards molecular and phenotypic information of cereals. To facilitate the comprehension of the terms that will be used in the rest of the paper and that the reader may not be familiar with, we provide in this section a brief description of the domain and of the data sources integrated in the CEREALAB laboratory.

The main entities about molecular data are three:

- **Gene**: it is the unit of heredity in living organisms, which controls the physical development of the organism. An allele is any one of a number of viable

DNA codings of the same gene occupying a given locus (position) on a chromosome.

- **QTL**: a quantitative trait locus, it is a region of DNA that is associated with a particular trait. Though not necessarily genes themselves, QTLs are stretches of DNA that are closely linked to the genes that underlie the trait in question.
- **Marker**: it is a known DNA sequence (e.g. a gene or part of gene) that can be identified by a simple assay, associated with a certain phenotype. A genetic marker may be a short DNA sequence, such as a sequence surrounding a single base-pair change, or long one, like microsatellites.

All these entities have their own specific attributes, for example their chromosome, which is a physically organized piece of DNA that contains Genes or QTLs; or their allele, which is any one of a number of viable DNA codings that occupies a given locus (position) on a chromosome.

The term **Germplasm** identifies an assemblage of plants that has been selected for a particular attribute or combination of attributes and is clearly distinct, uniform and stable in its characteristics. The **Trait** is an inherited feature of a plant, and is thus influenced by genes and QTLs.

3.2 The Data Sources

The data sources that have been integrated are briefly described in this section. The web databases Gramene and Graingenes have been chosen as data sources for the molecular data, since they were indicated to be the most relevant regarding the species involved in the project, i.e. rice, barley and wheat. Both these sources provide a traditional web interface to obtain molecular data. Gramene provides the dump of its relational database to be downloaded and restored locally. Moreover, Gramene is the developer of the Cereal Plant Trait Ontology and it allows browsing in its web site this ontology, that is a controlled vocabulary and taxonomy of phenotypic traits. As no molecular data are related to the terms of the Cereal Plant Trait Ontology, it results to be incomplete for the purpose of the CEREALAB laboratory. These two data sources have been integrated with molecular data obtained from a systematic genotyping work performed by the research group of the CEREALAB laboratory.

Phenotypic evaluations can be found in the GRIN database, which provides quantitative evaluations of numerous traits for many germplasms. Other phenotypic data have been collected by the research group of the CEREALAB laboratory from specific literature for regional germplasms (Emilia Romagna Data, ER Data) and from the italian National Council of Research in Agriculture (CRA), creating a local repository of these data to be integrated in our ontology. All these data sources, if considered separately, present incomplete information for the purpose of the CEREALAB laboratory and are sometimes overlapping.

4 THE MOMIS INTEGRATION PROCESS

MOMIS performs information extraction and integration from both structured and semi-structured data sources in a semi-automatic way. In this case, all the data sources involved are relational databases, but the system can deal also with XML and XSD sources and other ontologies expressed in the standard W3C OWL language.

The GS realized with the MOMIS system is expressed using the ODL_I^3 language, an extension of the ODL language, an object-oriented language developed by ODMG. ODL_I^3 is transparently translated into a Description Logic (Beneventano, Bergamaschi, and Sartori, 2003; Bergamaschi et al., 2001), and allows representing in a common data model different kinds of data sources and the view resulting from the integration process.

The GS is composed of Global Classes. Each Global Class includes several Global Attributes. Moreover, the GS elements are annotated according to the WordNet lexical database (Bergamaschi et al., 2007b), which provides an easily understandable meaning for each GS element.

The MOMIS integration process for building the GS, shown in Figure 2, has five phases:

- **Local source schemata extraction.** Wrappers automatically extract sources schemas. Such schemas are then translated into the common language ODL_I^3.
- **Local source annotation with respect to WordNet.** The integration designer selects one or more meanings for each element of a local source schema, according to the WordNet lexical database. A tool supports the integration designer: some WordNet synsets are suggested for each source element. Annotation is semi-automatically performed (Bergamaschi et al., 2007b).
- **Common Thesaurus generation.** Starting from the annotated local schemas, MOMIS discovers relationships describing inter- and intra-schema knowledge about classes and attributes of the source schemata that are inserted in the Common Thesaurus. The Common Thesaurus is incrementally built starting from schema-derived relationships, i.e. automatically extracted intra-schema relationships from each schema separately. Then, the relationships existing in the WordNet database between the annotated meanings are exploited to generate relationships between the respective elements (classes, attributes), called lexicon-derived relationships. The Integration Designer may add new relationships to capture specific domain knowledge, and finally, by means of a Description Logics reasoner, ODB-Tools (Beneventano, Bergamaschi and Sartori, 2003) (which performs equivalence and subsumption computation), new relationships are inferred and the transitive closure is computed.
- **GS generation.** MOMIS exploits the relationships included in the Common Thesaurus to generate an affinity matrix showing the similarity measure of the elements of the sources. A hierarchical clustering

technique applied to this affinity matrix groups similar classes of different sources in clusters, then generating a global schema (GS) and sets of mappings with local schemata (Bergamaschi et al., 2001).

- **GS annotation.** Exploiting the annotated local schemata and the mappings between local and global schemata, the MOMIS system semi-automatically assigns name and meaning to each class of the global schema.

A more detailed description of the MOMIS integration process can be found in Beneventano et al. (2003) and in Bergamaschi and Sala (2006).

The GS obtained at the end of the integration process can be translated from ODL_I^3 and exported into the OWL-DL language. The GS is defined "virtual" since only the schemas of the different data sources are integrated, while the data (instances) reside in the local sources and are not stored in the global schema. In this way, every time a query is posed over the GS, it is translated into sub-queries that are performed locally on the data sources guaranteeing that the data obtained are always up to date. The query processing is briefly described in section 6.

5 THE CEREALAB ONTOLOGY

The GS obtained with MOMIS can be seen as an ontology of the underlying sources. This ontology allows correlating the molecular data of Gramene, Graingenes and the CEREALAB laboratory with the phenotypic data of the GRIN database and those collected by the CEREALAB laboratory.

In this way, molecular data about genes and QTLs and information about their associated molecular markers are available. For each gene and QTL it is possible to retrieve its associated germplasms, i.e. the cultivars where that gene/QTL has been identified. Genes and QTLs are also associated with traits, and phenotypic evaluations of each of these traits are available for many germplasms.

Part of the ontology can be seen in Figure 3.

The ontology is composed of 66 Classes and 916 Attributes and is divided in two parts or sub-ontologies: the first containing genotypic data, and the second one containing phenotypic data. Genotypic data emerge from the integration of the molecular data sources such as Gramene and Graingenes and are instances of the classes `Gene`, `QTL`, `Markers` and `Traits`. The `markers` can be `marker_for` instances of the classes `Gene` or `QTL`. One or more genes or QTLs can affect each trait.

Phenotypic data are divided into six categories chosen among those of major interest for the cereal breeders: Abiotic Stress, Biotic Stress, Growth and Development related traits, Quality traits and Yield traits. In Figure 3 only the `Biotic_Stress` class is reported for the sake of readability. For each trait the specific evaluation of a germplasm for that trait is available.

Genes and QTLs are related to phenotypic data indicating their presence in a germplasm for which a quantitative phenotypic evaluation is available.

Thanks to the combined information available in our ontology, it is possible to find the specific molecular markers that can identify genes or QTLs that express a particular phenotypic trait. In this way genotypic selection of cereals cultivars can be performed starting from phenotypic data.

6 QUERYING THE INTEGRATED ONTOLOGY

The MOMIS Query Manager allows the user to pose a query expressed in the SQL language over the ontology and to obtain a unified answer from all the data sources integrated in the GS (see Beneventano and Bergamaschi (2007) for a technical description).

When the MOMIS Query Manager receives a query, it rewrites the global query as an equivalent set of queries expressed on the local schemata (local queries); this query translation is carried out by considering the mappings between the GS and the local schemata. Since MOMIS follows a Global as View (GAV) approach (Lenzerini, 2002), where the contents of the mediated schema is expressed in terms of the local sources schemata, this mapping is expressed by specifying, for each global class C, a mapping query QC over the schemata of the local classes belonging to C. The system automatically generates the mapping query QC, by extending the Full Disjunction (FD) operator (Galindo-Legaria, 1994) and exploiting the Data Transformation Functions, which are defined by the user and represent the mapping of local attributes into the attributes of the GS. The query translation is thus performed by means of query unfolding, i.e. by expanding a global query on a global class C of the GS according to the definition of the mapping query QC. Results from the local sources are then merged exploiting reconciliation techniques proposed by Naumann, Freytag, and Leser (2004) and proposed to the user (Beneventano and Bergamaschi, 2007).

In order to assure full usability of the system to users who do not know the SQL language, a graphical user interface has been developed to compose queries over the GS.

This interface, shown in Figure 4, presents in a tree representation the ontology, showing ISA relationships among the classes. The user can select the global classes to be queried and their attributes are shown in the "Global Class Attributes" panel with a simple click. Then the attributes of interest can be selected, specifying, if necessary, a condition in the "Condition" panel with usual relational predicates and logic operators. More than one global class can be joined just choosing one of the "Referenced Classes" of the currently selected class with no need to specify any join condition between the classes as it is automatically inserted. The graphical query, including selections and conditions specified by the user, is then

automatically translated into an SQL query and sent to the MOMIS Query Manager to be executed.

Figure 4 shows an example of the formulation of the query "retrieve all the QTLs that affect the resistance of a plant to the fungus "Fusarium" ". The user wants to find which QTLs, i.e. which pieces of DNA, influence the resistance of a plant to a particular fungus, "Fusarium" in this case, that can affect a plant with a disease and eventually cause its death. The result of the query, i.e. the QTLs that can express a high resistance to this fungus, allows the breeder to find a molecular marker that can help him to identify the presence of the QTL in the plant genome, and thus to decide whether to choose or not that germplasm for breeding. To do this, the user selects the class `QTL` from the tree on the left side representing the GS. All the attributes of QTL are shown in the tree in the middle panel. Then, the user adds to the selection the "Referenced Class" `Trait_affected_by_qtl`. All the attributes of this class are then automatically added to the "Global Class Attributes" panel, and the user may select attributes from this global class. To restrict the query only to the Fusarium-related QTLs, it is just needed to add in the "Condition" panel the condition:

```
Trait_affected like fusarium.
```

Then, clicking the button "Execute Query", the following query is composed, shown in the right side panel and sent to the MOMIS Query Manager:

```
SELECT Q.*, T.trait_affected
FROM Trait_affected_by_qtl as T, Qtl as Q
WHERE T.qtl_name=Q.name
AND T.trait_affected like '%fusarium%'
```

The result presented to the user is shown in Figure 5.

7 CONCLUSIONS AND FUTURE WORK

We created the CEREALAB ontology, which includes both molecular and phenotypic data about wheat, barley and rice, integrating existing molecular and phenotypic data sources and data provided by the CEREALAB laboratory. In this paper we presented this ontology, its creation process and the graphical user interface available to compose queries over it.

The CEREALAB ontology will improve the cereal breeding process as it allows cereal breeders to find the right molecular markers to be used to intentionally breed certain traits, or combinations of traits, over others. To this purpose, the combination of molecular data and phenotypic evaluation of traits is required. No resource was available so far that combined both these two kind of data, but it was necessary to separately access different sources and then manually combine partial results. Instead, with our work both molecular and phenotypic data are available through a single query by means of an easy-to-use graphical interface. The CEREALAB ontology thus overcomes the Cereal Plant Trait Ontology as it combines molecular and phenotypic data and associates quantitative evaluations of the phenotypic traits of the TO with molecular data. The advantage is evident, since all the information required are available through a single interface with no need to navigate through several websites, access different data sources and combine results manually. Anyway future work is still needed to provide keyword based search capabilities for our virtual approach, since the experience of the users says that it is complimentary to the tree based exploration of the ontology. Sometimes, in fact, when the user has not clear in mind what he/she is looking for, a keyword-based search is more suitable to address the user to the part of the ontology to be queried. The problem is that in a virtual approach data are not available at the mediator level and it is thus not possible to build indexes like in traditional information retrieval keyword-based search approaches. As a starting point, in Bergamaschi et al. (2007a), a method for extracting a synthesized view of an attribute's values has been proposed that can be exploited by the user when creating or refining a search query in data integration systems.

ACKNOWLEDGEMENT

The work presented in this paper has been supported by the Emilia Romagna P.R.R.I.I.T.T. through the CEREALAB laboratory (Non-GM Biotechnologies Laboratory for the Seed Industry) and the SITEIA laboratory (Agro-Food Safety Technology and Innovation).

REFERENCES

Beneventano, D., Bergamaschi, S. (2007): Semantic Search Engines based on Data Integration Systems. In: *Semantic Web Services: Theory, Tools and Applications.* Idea Group Publishing

Beneventano, D., Bergamaschi, S., Guerra, F., Vincini, M. (2003): Synthesizing an integrated ontology. *IEEE Internet Computing* 7(5) 42–51

Beneventano, D., Bergamaschi, S., Sartori, C. (2003): Description logics for semantic query optimization in object-oriented database systems. *ACM Trans. Database Syst.* 28 1–50

Bergamaschi, S., Castano, S., Vincini, M. (1999): Semantic integration of semistructured and structured data sources. *SIGMOD Record* 28(1) 54–59

Bergamaschi, S., Castano, S., Vincini, M., Beneventano, D. (2001): Semantic integration of heterogeneous information sources. *Data Knowl. Eng.* 36(3) 215–249

Bergamaschi S., Guerra F., Orsini M., Sartori C. (2007a): Extracting Relevant Attribute Values for Improved Search, *IEEE Internet Computing*, vol. 11, no. 5, pp. 26-35, Sept/Oct (special issue on Semantic-Web-Based Knowledge Management)

Bergamaschi, S., Po, L., Sala, A., Sorrentino, S. (2007b): Automatic annotation for p2p data integration systems: the wordnet domains disambiguation approach. In: *Fifth International Workshop on Databases, Information Systems and Peer-to-Peer Computing (DBISP2P 2007)* at *VLDB 2007 33st International Conference on Very Large Data Bases, September 24, 2007*, University of Vienna, Austria.

Bergamaschi, S., Sala, A. (2006): Virtual integration of existing web databases for the genotypic selection of cereal cultivars. In Meersman, R., Tari, Z., eds.: *OTM Conferences (1). Volume 4275 of Lecture Notes in Computer Science.* Springer 909–926

Davidson, S.B., Overton, G.C., Tannen, V., Wong, L. (1997): Biokleisli: A digital library for biomedical researchers. *Int. J. on Digital Libraries* 1(1) 36–53

Davidson, S.B., Crabtree, J., Brunk, B.P., Schug, J., Tannen, V., Overton, G.C., Jr., C.J.S. (2001): K2/kleisli and gus: Experiments in integrated access to genomic data sources. *IBM Systems Journal* 40(2) 512–531

Galindo-Legaria, C.A. (1994): Outerjoins as disjunctions. In Snodgrass, R.T., Winslett, M., eds.: *SIGMOD Conference*, ACM Press 348–358

Haas, L.M., Schwarz, P.M., Kodali, P., Kotlar, E., Rice, J.E., Swope, W.C. (2001): Discoverylink: A system for integrated access to life sciences data sources. *IBM Systems Journal* 40(2) 489–511

Halevy A.Y., Rajaraman A., and Ordille J. J. (2006): Data integration: The teenage years. In Dayal U., Whang K.-Y., Lomet D. B., Alonso G., Lohman G. M., Kersten M. L., Cha S. K., and Kim Y.-K., editors, *VLDB*, pages 9–16. ACM, 2006

Hernandez, T., Kambhampati, S. (2004): Integration of biological sources: Current systems and challenges ahead. *SIGMOD Record* 33(3) 51–60

Köhler J., Philippi S., Lange M. (2003): SEMEDA: ontology based semantic integration of biological databases. *Bioinformatics*, Dec 12;19(18):2420-7

Lenzerini M. (2002): Data integration: A theoretical perspective. In L. Popa, editor, *PODS*, pages 233–246, ACM.

Naumann, F., Freytag, J.C. and Leser, U.(2004): Completeness of integrated information sources. *Inf. Syst.*, 29(7):583–615

Stevens, R., Baker, P.G., Bechhofer, S., Ng, G., Jacoby, A., Paton, N.W., Goble, C.A., Brass, A. (2000): Tambis: Transparent access to multiple bioinformatics information sources. *Bioinformatics* 16(2) 184–186

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

WEBSITES

The Database Group at the University of Modena and Reggio Emilia,
http://www.dbgroup.unimo.it

CEREALAB laboratory website,
http://www.cerealab.org

The Cereal Plant Trait Ontology,
http://www.gramene.org/plant_ontology

Germplasm Resource Information Network, GRIN,
http://www.ars-grin.gov/

Graingenes,
http://wheat.pw.usda.gov/GG2

Gramene,
http://www.gramene.org

Italian Council of Research in Agriculture, CRA,
http://www.entecra.it/

The MOMIS system,
http://www.dbgroup.unimo.it/Momis

Object Data Management Group,
http://www.odmg.org

The Open Biomedical Ontologies Foundry (OBO)
http://www.obofoundry.org/

Web Ontology Language, OWL
http://www.w3.org/2004/OWL/

Wordnet,
http://wordnet.princeton.edu/

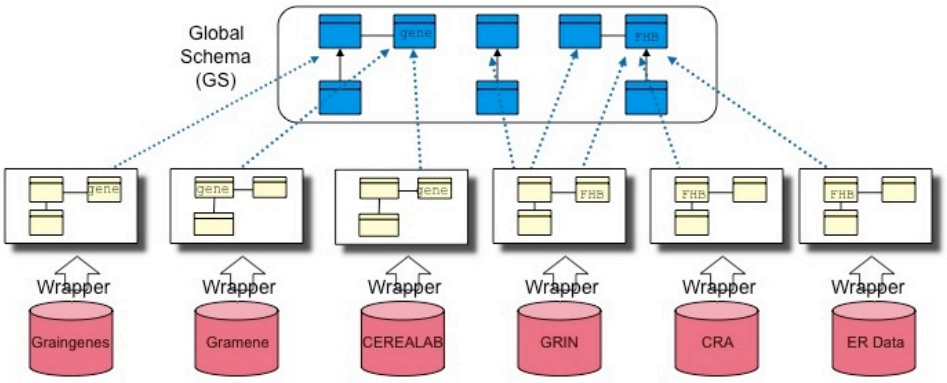

Figure 1 Creating the GS with the MOMIS System

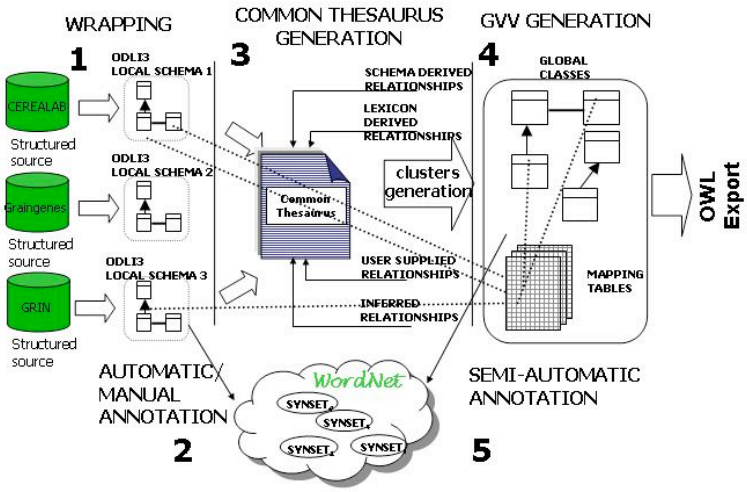

Figure 2 Integration Process Overview

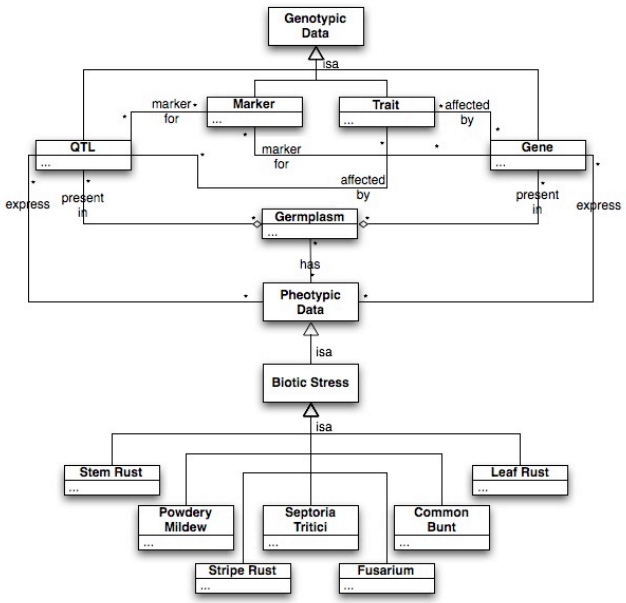

Figure 3 UML Class Diagram representing an excerpt of the Ontology

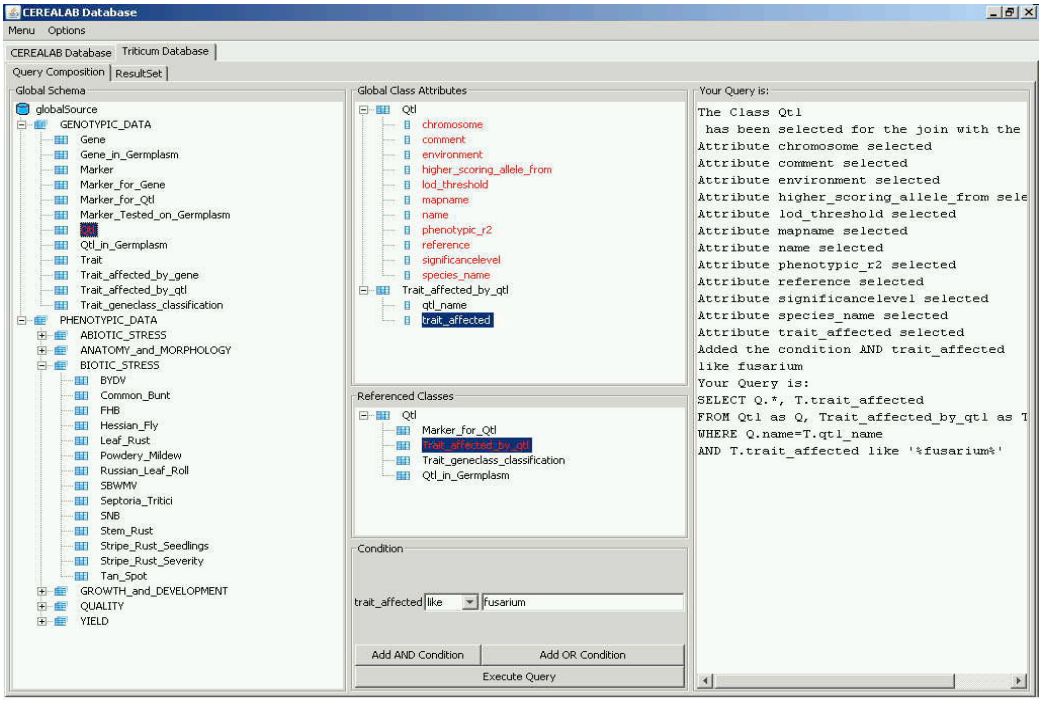

Figure 4 The Graphical User Interface the composition of the query: "Retrieve all the QTLs that affect the resistance of a plant to the fungus "Fusarium""

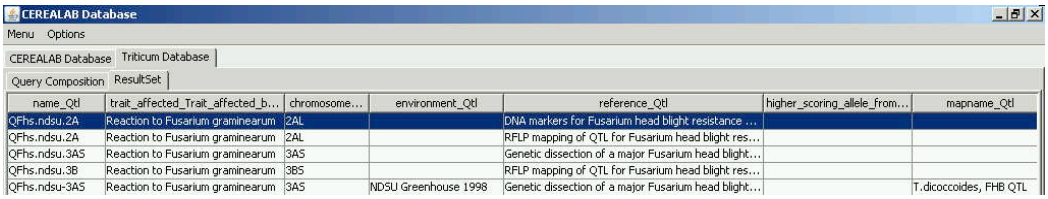

Figure 5 The ResultSet of the query: "Retrieve all the QTLs that affect the resistance of a plant to the fungus "Fusarium""