Creating and Querying an Integrated Ontology for Molecular and Phenotypic Cereals Data *

Sonia Bergamaschi and Antonio Sala

Dipartimento di Ingegneria dell'Informazione
Universitá degli Studi di Modena e Reggio Emilia
bergamaschi.sonia@unimore.it, sala.antonio@unimore.it

Abstract. In this paper we describe the development of an ontology of molecular and phenotypic cereals data, realized by integrating existing public web databases with the database developed by the research group of the CEREALAB project[1]. This integration is obtained using the MOMIS system (Mediator envirOnment for Multiple Information Sources), a mediator based data integration system developed by the Database Group of the University of Modena and Reggio Emilia[2]. MOMIS performs information extraction and integration from both structured and semi-structured data sources in a semi-automatic way. Information integration is performed in a semi-automatic way, by exploiting the knowledge in a Common Thesaurus (defined by the framework) and the descriptions of source schemas with a combination of clustering and Description Logics techniques. The result of the integration process is a Global Virtual Schema (GVV) of the underlying data sources for which mapping rules and integrity constraints are specified to handle heterogeneity. Each GVV element is annotated w.r.t. the WordNet lexical database[3]. The GVV can be queried transparently with regards to integrated data sources using an easy to use graphical interface regardless of the specific languages of the source databases.

1 Introduction and Motivation

In the last few years numerous public data sources have been realized and are now 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. For all the above problems, sometimes a simple information search can take long time and eventually fails, even because of the number of different data sources to be

* The work presented in this paper was partially supported by MUR FIRB NeP4B - Network Peer for Business project (http://www.dbgroup.unimo.it/nep4b) and by the IST FP6 STREP project 2006 STASIS (http://www.dbgroup.unimo.it/stasis).

[1] http://www.cerealab.org

[2] http://www.dbgroup.unimo.it

[3] http://wordnet.princeton.edu/

accessed. What is needed by users is, thus, to have access to the information available in different data sources in a transparent and easy way, independently from the format of the different sources.

There are different public reference databases regarding cereals molecular data: Graingenes[4], for wheat and barley, and Gramene[5] for rice. These databases present also descriptions of phenotypic characters, but no quantitative evaluation of such traits is available. On the other hand, the American Germplasm Resources Information Network (GRIN)[6] provides phenotypic information about many germplasms, but no molecular data.

The aim of our work is thus to create a unique ontology with a global interface, that integrates the above mentioned public data sources providing both molecular and phenotypic data about wheat, barley and rice. Moreover, the ontology has to easily integrate new molecular data coming from the research activity of the CEREALAB project.

The integration process of the public databases and the CEREALAB database is performed with the MOMIS system[7](for further details on the integration process, see [1–3]).

The work presented in this paper has been conducted as a joint collaboration between the DBGroup and the Agrarian faculty of the University of Modena and Reggio Emilia within the CEREALAB project. As far as we know, no resource is available containing both these two kinds of data for the purpose of this project. For this reason, we developed a Global Virtual View (GVV) which is the integration of existing molecular and phenotypic data sources with data provided by the CEREALAB project. The GVV can be seen as an ontology of the underlying sources.

Other ontologies about these domain exist, but none of these correlates phenotypic data with molecular data. For example the Trait Ontology (TO)[8] 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.

Our ontology overcomes the TO as it integrates the trait ontology with molecular data related to phenotypic data.

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 IT expertise, it follows that the usage of the system has to be as much user-friendly as possible, and it is necessary to provide the users with a graphical interface to query this ontology.

[4] http://wheat.pw.usda.gov/GG2
[5] http://www.gramene.org
[6] http://www.ars-grin.gov/
[7] http://www.dbgroup.unimo.it/Momis
[8] http://www.gramene.org/plant_ontology/

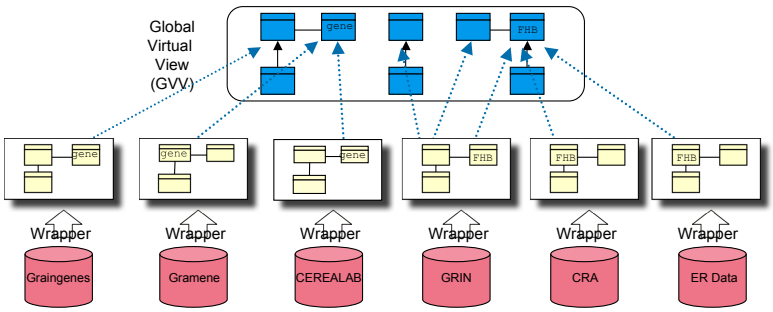

Fig. 1. Creating the GVV with the MOMIS System

The goal of this work is to present the ontology with its interface and sketch the translation of graphical queries into queries executable by the MOMIS system.

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

2 Description of the Domain and of the Data sources

To facilitate the comprehension of the terms involved in our project, in this section we provide a brief description of the domain of the CEREALAB project and of the data sources integrated. 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, such as its chromosome, which is physically organized piece of DNA that contains Genes or QTLs; or its 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.

The web databases Gramene and Graingenes have been chosen as data sources for the molecular data as 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 form to obtain molecular data.

Moreover, Gramene is the developer of the Trait Ontology and it allows to browse this ontology, which is only a controlled vocabulary and a taxonomy of phenotypic traits. As no molecular data are related to the terms of the TO, it results to be incomplete for the purpose of the CEREALAB project.

These two data sources have been integrated with molecular data obtained from a systematic genotyping work performed by the CEREALAB research group.

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 CEREALAB research group 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 project and are sometimes overlapping.

3 The Momis Integration Process

MOMIS performs information extraction and integration from both structured and semistructured data sources. In this case, all the data sources involved are relational databases, but the system can deal also with XML and XSD sources and existing ontologies expressed in OWL. The GVV 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[9]. ODL_{I^3} is transparently translated into a Description Logic [4, 5, 2]. ODL_{I^3} allows to represent in a common data model different kinds of data sources and the view resulting from the integration process. The GVV is composed of Global Classes. Each Global Class includes several Global Attributes. Moreover, the GVV elements are annotated according to the WordNet lexical reference system[10], which provides an easily understandable meaning for each GVV element.

[9] `http://www.odmg.org/`

[10] `http://wordnet.princeton.edu/`

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

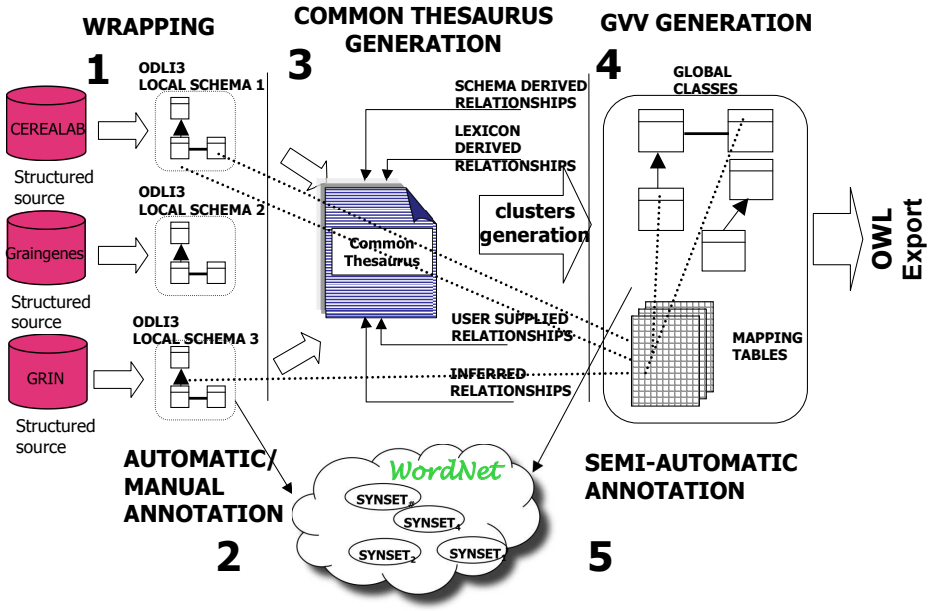

Fig. 2. Integration Process Overview

1. **Local source schemata extraction**. Wrappers automatically extract sources schemas. Such schemas are then translated into the common language ODL_{I^3}.
2. **Local source annotation with WordNet**. The integration designer selects a meaning for each element of a local source schema, according to the WordNet lexical ontology. A tool supports the integration designer: some WordNet synsets are suggested for each source element. Annotation is semi-automatically performed [6, 7].
3. **Common thesaurus generation**. Starting from the annotated local schemas, MOMIS extracts 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, 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 [8](which performs equivalence

and subsumption computation), infers new relationships and computes the transitive closure of Common Thesaurus relationships.
4. **GVV 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 elements of different sources in clusters, then generating a global schema (GVV) and sets of mappings with local schemata [2].
5. **GVV 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 element of the global schema.

The GVV obtained at the end of the integration process can be translated and exported in the OWL language.

A more detailed description of the MOMIS integration process can be found in [9, 3].

4 The integrated Ontology

The GVV obtained with MOMIS can be seen as an ontology of the underlying sources. This ontology allows to correlate the molecular data of Gramene, Graingenes and the CEREALAB project with the phenotypic data of the GRIN database and those collected by the CEREALAB project. 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 trait are available for many germplasms. Part of the ontology can be seen in Fig.3.

The ontology is thus divided in two parts: the first containing genotypic data, and the second one containing phenotypic data. Genotypic data are divided into the classes `Gene`, `QTL`, `Markers` and `Traits`. The markers can be `marker_for` instances of the classes `Gene` or `QTL`. Each trait can be affected by one or more genes or QTLs.

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 Fig.3 only the `Biotic_Stress` class is reported for the sake of readability. For each trait the specific value 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.

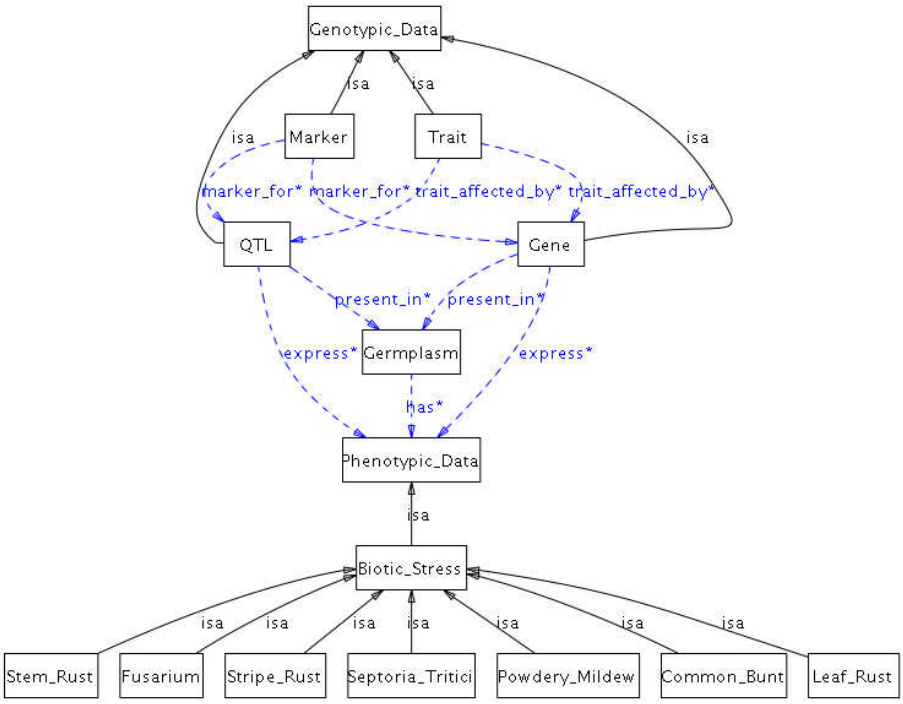

Fig. 3. An excerpt of the Integrated Ontology visualized with the Ontoviz Plugin for Protégé

5 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 GVV (see [10] 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 mapping between the GVV and the local schemata. Since MOMIS follows a Global as View (GAV) approach, where the contents of the mediated schema is expressed in terms of queries over the sources, 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 [11] and exploiting the Data Transformation Functions, which are defined by the user and represent the mapping of local attributes into the attributes of the GVV. 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 GVV according to the definition of the mapping query QC. Results

from the local sources are then merged exploiting reconciliation techniques and proposed to the user [10].

In order to assure full usability of the system even to users who do not know the SQL language, a graphical user interface has been developed to compose queries over the GVV. This interface, shown in Fig.4, presents in a tree repre-

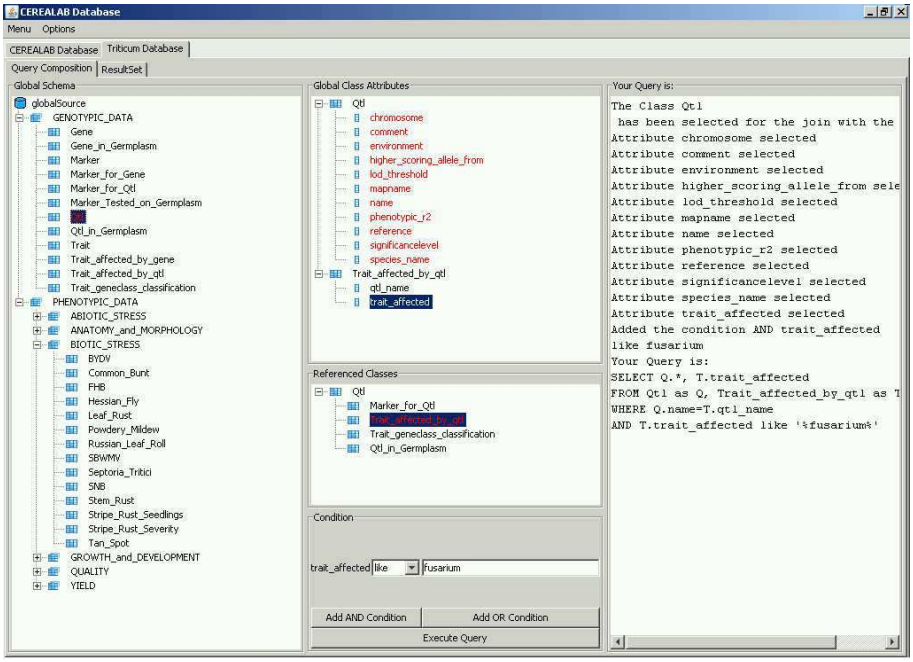

Fig. 4. The Graphical User Interface for querying the Integrated Ontology

sentation 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 the usual SQL 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 among classes as it is automatically inserted. Selections and conditions specified by the user are then automatically translated into an SQL query and sent to the MOMIS Query Manager.

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"". This query lets the user 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 GVV. 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 Fig.5

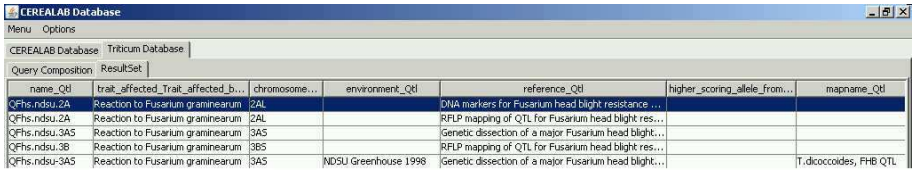

Fig. 5. The Result Set obtained querying the Ontology

6 Related Work

In the last few years the problem of data integration for biology has become really important both due to continuous increases in data volumes and the growing diversity in types of data that need to be managed. For example the Transparent Access to Multiple Bioinformatics Information Sources project, known as TAMBIS [12], 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. TAMBIS can seem similar to our approach but in this system mappings among the global schema and the local sources are constructed manually, while in MOMIS clusters of similar classes and mappings of global schema

classes with local schemas are automatically generated once the sources have been semi-automatically annotated. The process of generation of the GVV is thus semi-automatic.

BioKleisli [13] 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. Furthermore, no optimization based on source characteristics or source content is performed. K2 [14] 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 [15] 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 hide the sources and do not offer schema or data reconciliation.

A survey of these and some other well-known systems that are currently available can be found in [16].

As it can be seen, the data integration problem for biology has been addressed in numerous way, but as far as we know the approach presented in this paper is the first one that combines molecular and phenotypic data in an integrated ontology. All the other systems integrates only molecular data sources, while our combines molecular and phenotypic data. Moreover, except TAMBIS, usually the existing systems uses the SQL language to formulate queries, while in our system we developed a graphical interface for query formulation which is considered a necessity as the users of this kind of systems have low IT expertise and thus need a user-friendly system.

7 Conclusions

We created a unique ontology providing both molecular and phenotypic data about wheat, barley and rice, integrating existing molecular and phenotypic data sources and data provided by the CEREALAB project. In this paper we presented this ontology and the graphical user interface available to compose queries over the integrated ontology. The main advantage of our system is that retrieving data coming from numerous data sources requires only the use of a single interface instead of navigating through numerous web databases, querying them and then manually fusing the information obtained.

This integrated ontology can improve the breeding process as it allows cereal breeders to find the right molecular markers to be used to intentionally breeds certain traits, or combinations of traits, over others. To do this, access both to molecular data and phenotypic evaluation of traits is required. No resource was available so far that combined both these two kind of data and thus many data sources had to be accessed and the information obtained had to be combined manually. With our system both molecular and phenotypic data are available through a single graphical interface. Our integrated ontology thus overcomes the Trait Ontology as it combines molecular and phenotypic data and associates quantitative evaluations of the phenotypic traits of the TO with molecular data.

References

1. Bergamaschi, S., Castano, S., Vincini, M.: Semantic integration of semistructured and structured data sources. SIGMOD Record **28**(1) (1999) 54–59
2. Bergamaschi, S., Castano, S., Vincini, M., Beneventano, D.: Semantic integration of heterogeneous information sources. Data Knowl. Eng. **36**(3) (2001) 215–249
3. Bergamaschi, S., Sala, A.: 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 (2006) 909–926
4. Beneventano, D., Bergamaschi, S., Sartori, C.: Description logics for semantic query optimization in object-oriented database systems. ACM Trans. Database Syst. **28** (2003) 1–50
5. Beneventano, D., Bergamaschi, S., Lodi, S., Sartori, C.: Consistency checking in complex object database schemata with integrity constraints. IEEE Trans. Knowl. Data Eng. **10**(4) (1998) 576–598
6. Bergamaschi, S., Po, L., Sorrentino, S.: Automatic annotation for mapping discovery in data integration systems. In Meersman, R., Tari, Z., eds.: OTM Conferences (1). Lecture Notes in Computer Science, Springer (2007)
7. Bergamaschi, S., Po, L., Sala, A., Sorrentino, S.: 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) to be held at VLDB 2007 33st International Conference on Very Large Data Bases. University of Vienna, Austria, September 24, 2007
8. Bergamaschi, S., Beneventano, D., Sartori, C., Vincini, M.: Odb-qoptimizer: A tool for semantic query optimization in oodb. In Gray, W.A., Larson, P.Å., eds.: ICDE, IEEE Computer Society (1997) 578
9. Beneventano, D., Bergamaschi, S., Guerra, F., Vincini, M.: Synthesizing an integrated ontology. IEEE Internet Computing **7**(5) (2003) 42–51
10. Beneventano, D., Bergamaschi, S.: Semantic Search Engines based on Data Integration Systems. In: Semantic Web Services: Theory, Tools and Applications. Idea Group Publishing (2007)
11. Galindo-Legaria, C.A.: Outerjoins as disjunctions. In Snodgrass, R.T., Winslett, M., eds.: SIGMOD Conference, ACM Press (1994) 348–358
12. Stevens, R., Baker, P.G., Bechhofer, S., Ng, G., Jacoby, A., Paton, N.W., Goble, C.A., Brass, A.: Tambis: Transparent access to multiple bioinformatics information sources. Bioinformatics **16**(2) (2000) 184–186

13. Davidson, S.B., Overton, G.C., Tannen, V., Wong, L.: Biokleisli: A digital library for biomedical researchers. Int. J. on Digital Libraries **1**(1) (1997) 36–53
14. Davidson, S.B., Crabtree, J., Brunk, B.P., Schug, J., Tannen, V., Overton, G.C., Jr., C.J.S.: K2/kleisli and gus: Experiments in integrated access to genomic data sources. IBM Systems Journal **40**(2) (2001) 512–531
15. Haas, L.M., Schwarz, P.M., Kodali, P., Kotlar, E., Rice, J.E., Swope, W.C.: Discoverylink: A system for integrated access to life sciences data sources. IBM Systems Journal **40**(2) (2001) 489–511
16. Hernandez, T., Kambhampati, S.: Integration of biological sources: Current systems and challenges ahead. SIGMOD Record **33**(3) (2004) 51–60