Ontology-Based Mediation in the Amine System Project

Pisa June 2007

Prof. Dr. José F. Aldana Montes (jfam@lcc.uma.es)
Prof. Dr. Francisca Sánchez-Jiménez
Ismael Navas Delgado
Raúl Montañez
Almudena Pino-Ángeles
Aurelio A. Moya-García
José Luis Urdiales

http://asp.uma.es
Outline

- Family of problems to be solved
- Proposed solution: from Semantics to Data Integration
  - Semantic Directories
  - Ontology-Based Mediator
  - Specific Problem: Use Case
- Conclusions
Outline

- Family of problems to be solved
- Proposed solution: from Semantics to Data Integration
  - Semantic Directories
  - Ontology-Based Mediator
  - Specific Problem: Use Case
- Conclusions
- Systems biology is the study of an organism, viewed as an integrated and interacting network of genes, proteins and biochemical reactions which give rise to life. (Institute of Systems Biology)

- Instead of focusing on individual parts, the focus is on a complete system

→

**Need of integrated access**

to different data sources  
to enable the study of the system
- Family of problems to be solved
- **Proposed solution: from Semantics to Data Integration**
  - Semantic Directories
  - Ontology-Based Mediator
  - Specific Problem: Use Case
- Conclusions
Proposed Solution

- Requirements:
  - Easily extensible with new resources
  - Reusable elements
  - Possibility of developing different kind of applications (not only data integration)

- Decisions:
  - Take advantage of the Semantic Web
  - Annotate data sources with respect to ontologies
  - Reuse previous works
Proposed Solution

- Requirements:
  - Easily extensible with new databases
  - Reusable elements
  - Possibility of developing different kind of applications (not only data integration)

- Decisions:
  - Take advantage of the Semantic Web
  - Annotate data sources with respect to ontologies
  - Reuse previous works
Outline

- Family of problems to be solved
- Proposed solution: from Semantics to Data Integration
  - Semantic Directories
  - Ontology-Based Mediator
  - Specific Problem: Use Case
- Conclusions
Semantic Directories

Semantic Register

Resource Metadata Repository
Ontology Metadata Repository

Semantics

MM
OM
RM
OMV (ONTOLOGY METADATA VOCABULARY) is described with OWL, and each instance of the class OntologyImplementation represents an ontology registered in the Semantic Directory. It is possible to describe some relationships between ontologies.
SDMO

- OMV: link resources with registered ontologies.
- Resource: store information (query capabilities, schema, query interface, name and URI) about resources.
- Mapping: set the relationships between resources and ontologies. Each mapping is related with a similarity instance that establishes the similarity between ontology concepts and resource elements. The mapping class is related with OMV, Resource and Similarity class.
- Similarity: contains three properties (concept1, concept2 and similarityValue) to establish the similarity between an ontology concept and a resource element.
- User: deal with users in the applications.
Our goal is to provide applications which will make the semantics of the resources explicit through their commitment with an ontology registered in the Semantic Directory. The applications that can be developed using the Semantic Directory components depend on the extension of the infrastructure by means of new components (built on top of the Semantic Directory).
SD Conclusions

- Generic Infrastructure

- Basic Functionality

- Extended Functionalities requires Core Extensions: new metadata, interfaces, ...

- Fully implemented:
  - V 0.9.A: Java, BD MySQL, Racer (Concepts Classification), Web Services
  - V 0.9.B: Java, Metadata Files, Jena, Web Services
  - V 0.9.C: Java, Metadata Files, Jena, Corba CCM
Outline

- Family of problems to be solved
- Proposed solution: from Semantics to Data Integration
  - Semantic Directories
  - Ontology-Based Mediator
- Specific Problem: Use Case
- Conclusions
Mediator (Data Integration)

Discover available data sources

Select data sources & tools

Interact with each data source

Combine results

Users must locate data sources that are available online and can solve their problems

Users must select which data sources include relevant information for themselves

Users manually interact with each data source getting partial results

Users combine partial results to get a partial solution of the problem

Building the full solution
Approach

- GAV (BioBroker):
  + Easy query rewriting
  - Extension $\rightarrow$ Global Schema Changes

- LAV
  + Easy addition of new data sources
  - Complex Rewriting Process $\rightarrow$ simpler components will allow partial improvements
Approach

- GAV (BioBroker):
  + Easy query rewriting
  - Extension → Global Schema Changes

- LAV
  + Easy addition of new data sources
  - Complex Rewriting Process → simpler components will allow partial improvements
Main Characteristics

User Query, Ontology (Q,O)

Result (ontology instances)

Controller

Query Planner

Query Plan (QP)

Reasoner Component

Semantic Register

Resource Semantic Descriptions

Technical Users or Software Developers

Wrappers developed as Web Services

Data Service

Data Service

Data Service

Data Service

Technical Users or Software Developers

Resource Semantic Descriptions
Component Division

User Query, Ontology (Q,O)

Result (ontology instances)

Mappings Search

Resource Search

Ontology Search/reasoning

Data Service

Data Service

Data Service

Data Service

Mediator
Partial Improvements

User Query, Ontology (Q,O)

Result (ontology instances)

Q, O, Result (Ontology instances)

Q, O, Query Plan (QP)

Q, O, Query Plan (QP)

Q, O, Query Plan (QP)

Q, O, Query Plan (QP)

Q, O, Query Plan (QP)

Q, O, Query Plan (QP)

Q, O, Query Plan (QP)

Controller

Mapping Search

Ontology Search /reasoning

Resource Search

Data Service

Data Service

Data Service

Data Service

Query Planner

Controller

Integrator

Data Service

Data Service

Data Service

Data Service

Query Plan Solver/ Evaluator

Ontology Search /reasoning

Partial Improvements
Reusing elements and sharing semantics enable the mediator to find a query plan (QP) for the user query. This plan is then executed by interacting with the user interface to perform the corresponding call to the data services involved in the sub-queries. Results from data services (R1, ..., Rn) are composed.
Mediator Conclusions

- LAV
- Ontology Based
- Enabled partial improvements
- Limited Reasoning
- First beta version implemented (testing phase):
  - Test 1: Bioinformatics Resources
  - Test 2: Second Hand Car Resources
- Family of problems to be solved
- Proposed solution: from Semantics to Data Integration
  - Semantic Directories
  - Ontology-Based Mediator
  - Specific Problem: Use Case
- Conclusions
A common and useful strategy to find the 3D structure of a protein, which cannot be obtained by its crystallization, is to apply comparative modeling techniques. These work from the primary sequence of the unsolved protein and predict its 3D structure by comparing it to those of solved homologous proteins.
Domain Ontology

PubChem
Kegg
Brenda
Prosite

SWISS-Prot
PDB
BlastP
Modeller
Jmol
Next Step

User Query, Ontology (Q,O) → Controller

Result (ontology instances) → Q, O, Query Plan (QP) → Controller

Q, O, Result (Ontology Instances) → Q, O, Query Plan (QP) → Controller

Q, O, Query Plan (QP) → QP, {R1, ..., Rn} → Controller

Controller → QP, {R1, ..., Rn} → Resource Search/Reasoning

Resource Search/Reasoning → Query Planner

Query Planner → Mappings Search

Mappings Search → Ontology Search/Reasoning

Ontology Search/Reasoning → Data Service

Data Service → Data Service

Data Service → Data Service

Data Service → Integrator

Integrator
Application
Outline

- Family of problems to be solved
- Proposed solution: from Semantics to Data Integration
  - Semantic Directories
  - Ontology-Based Mediator
  - Specific Problem: Use Case
- Conclusions
Conclusions

- SD: Generic Infrastructure
- Two Ontologies to manage metadata
- SDMO: needs improvements and extensions
- Mediator
- Needs testing and improvement
- Study the addition of reasoning in the integration
Conclusions

Use Case: Protein structures contain fundamental information regarding their function, location and interactions, which is most of the information in their biological missions. Combining information integration with prediction techniques (as an automatic process) results in efficient information retrieval and expands the applicability spectrum of structural bioinformatics techniques to non experienced users.
Conclusions

- The problem presented is important in our context, as performing this process automatically will reduce the effort required to solve it.
- Genome Projects have exponentially increased the number of known polypeptide sequences. Thus, any effort to improve efficiency for the extraction of structural information at its highest level should help the advance of many ongoing Systems Biology projects.
Conclusions

The main limitation found is the maintenance of the data services, because the developed ones make use of public databases that are not under our control.

The long term success of this and similar proposals rely on the collaboration of data and tool owners.
Contact

José F. Aldana-Montes
jfam@lcc.uma.es

http://asp.uma.es