http://asp.uma.es

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

K·h·a·o·s INVESTIGACION

LENGUAJES Y CIENCIAS DE LA COMPUTACIÓN UNIVERSIDAD DE MÁLAGA

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

UNIVERSIDAD DE MÁLAGA

K·h·a·o·s INVESTIGACION

LENGUAJES Y CIENCIAS DE LA COMPUTACIÓN UNIVERSIDAD DE MÁLAGA

Family of Problems

- 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

Outline

- 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)

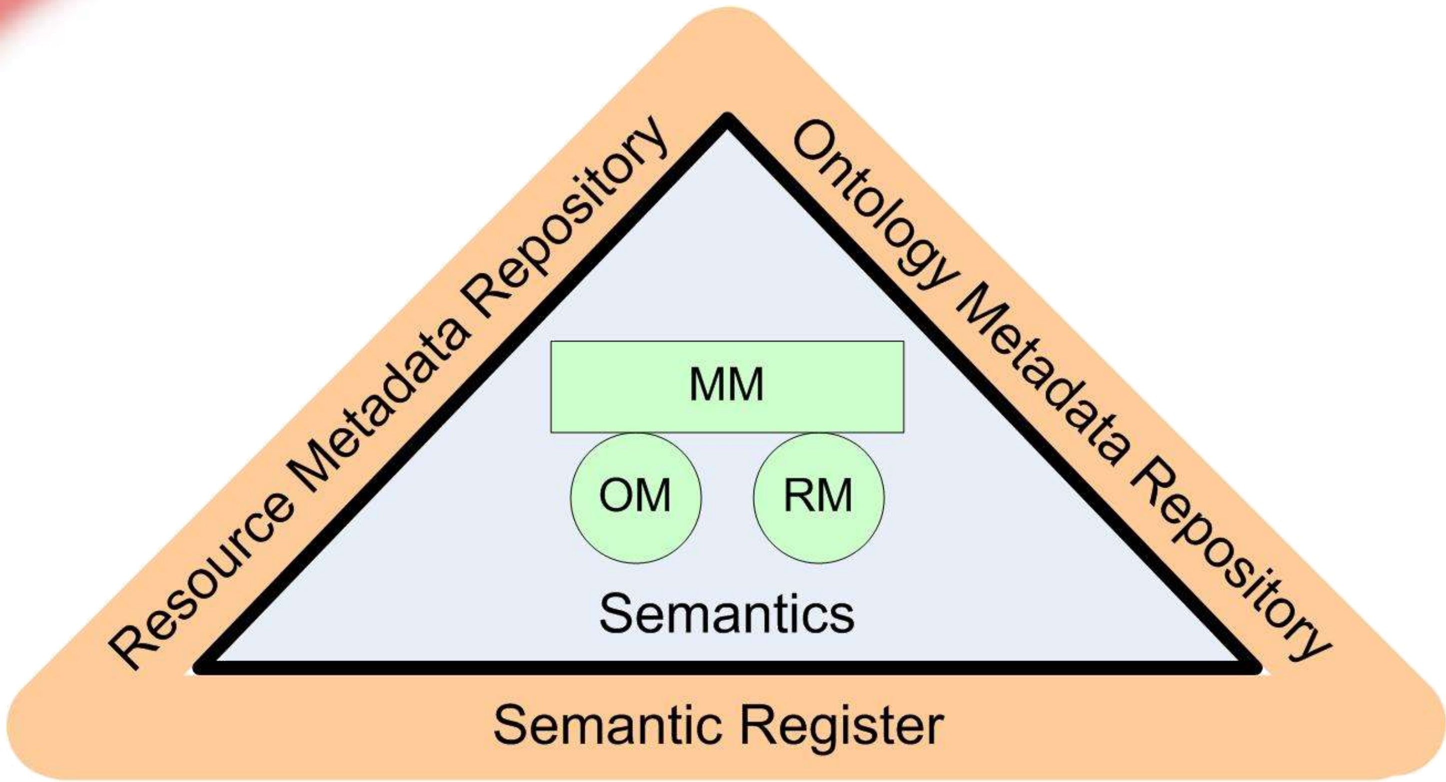
- Decisions:

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

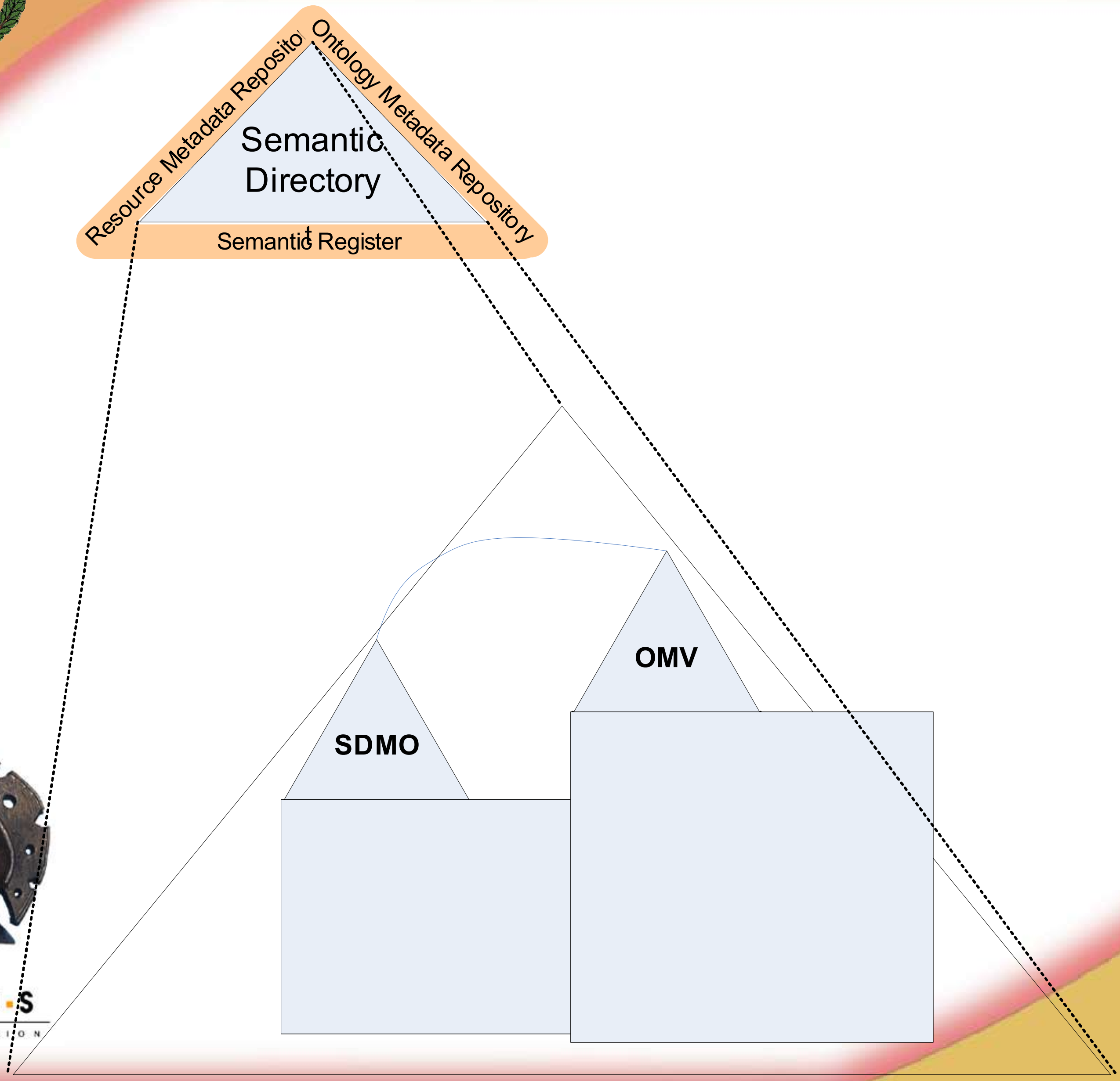
Outline

- 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

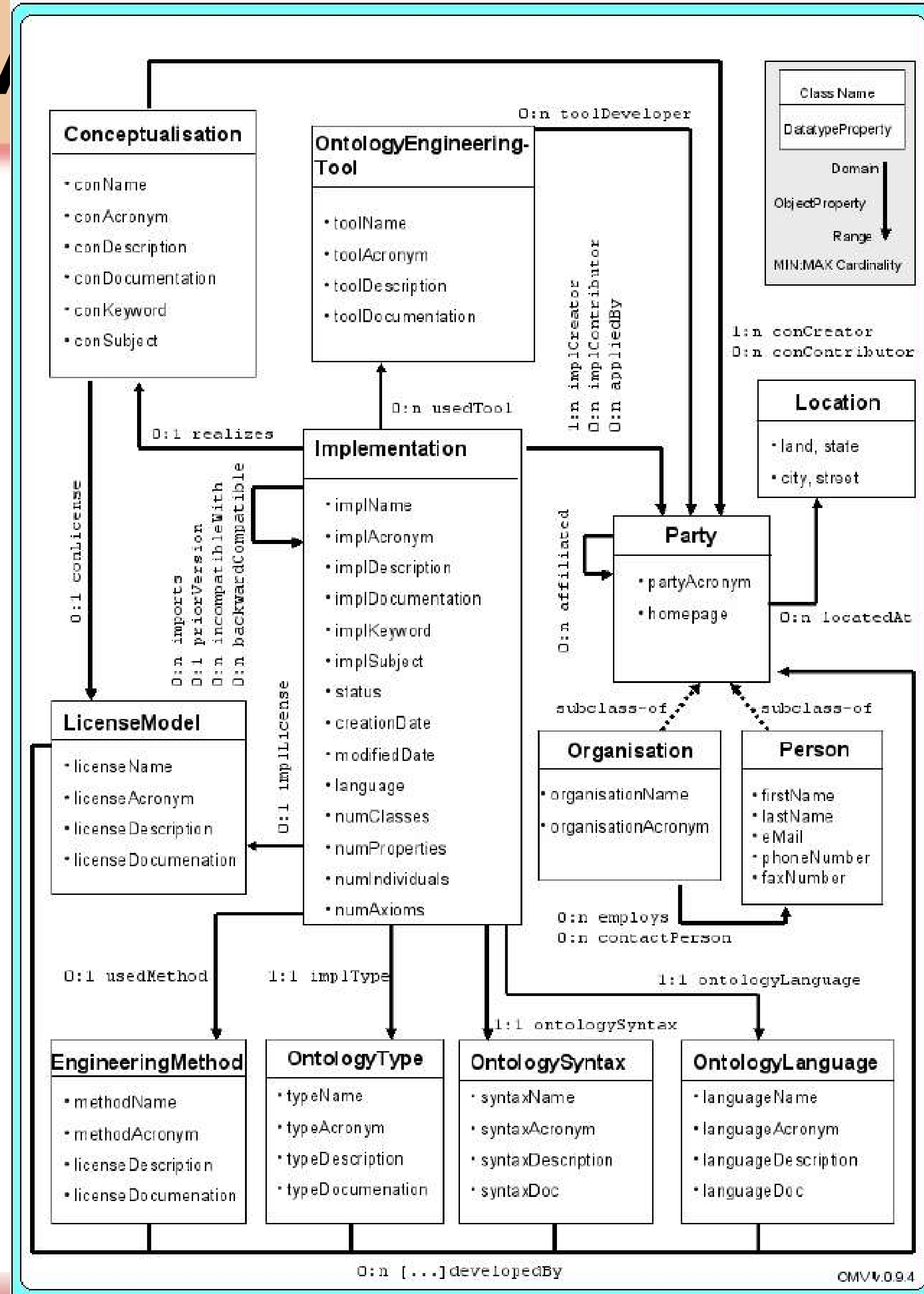


SD-Core Metadata

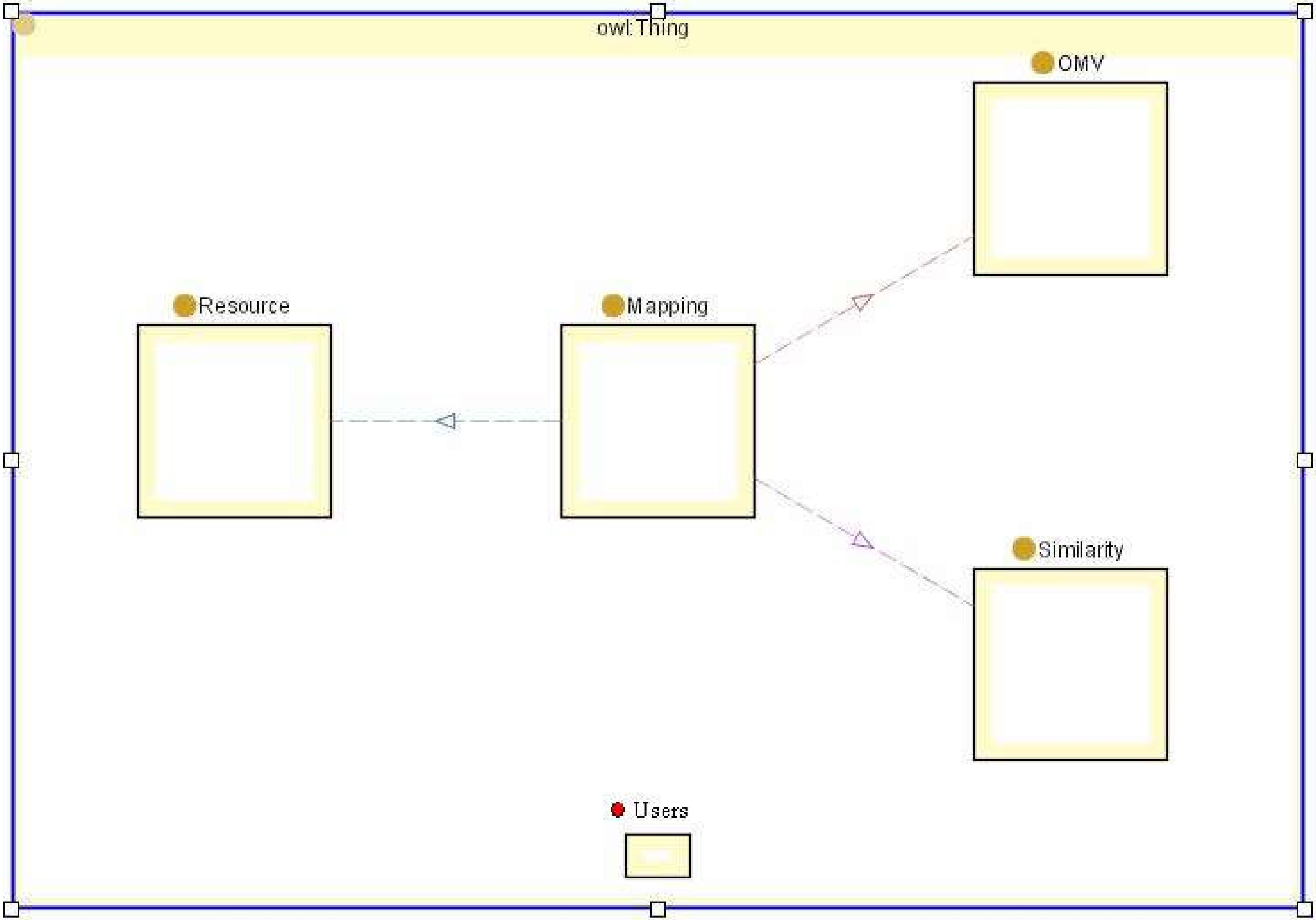


OM

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



Interfaces

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

K·h·a·o·s
INVESTIGACION

LENGUAJES Y CIENCIAS DE LA COMPUTACIÓN
UNIVERSIDAD DE MÁLAGA

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

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

Select data sources & tools

Users must select which data sources include relevant information for themselves

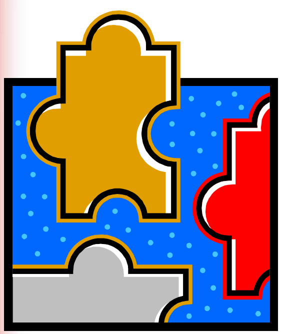
Interact with each data source

Users manually interact with each data source getting partial results

Combine results

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

Building the full solution



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

Approach

- GAV (BioBroker):

    + Easy query rewriting

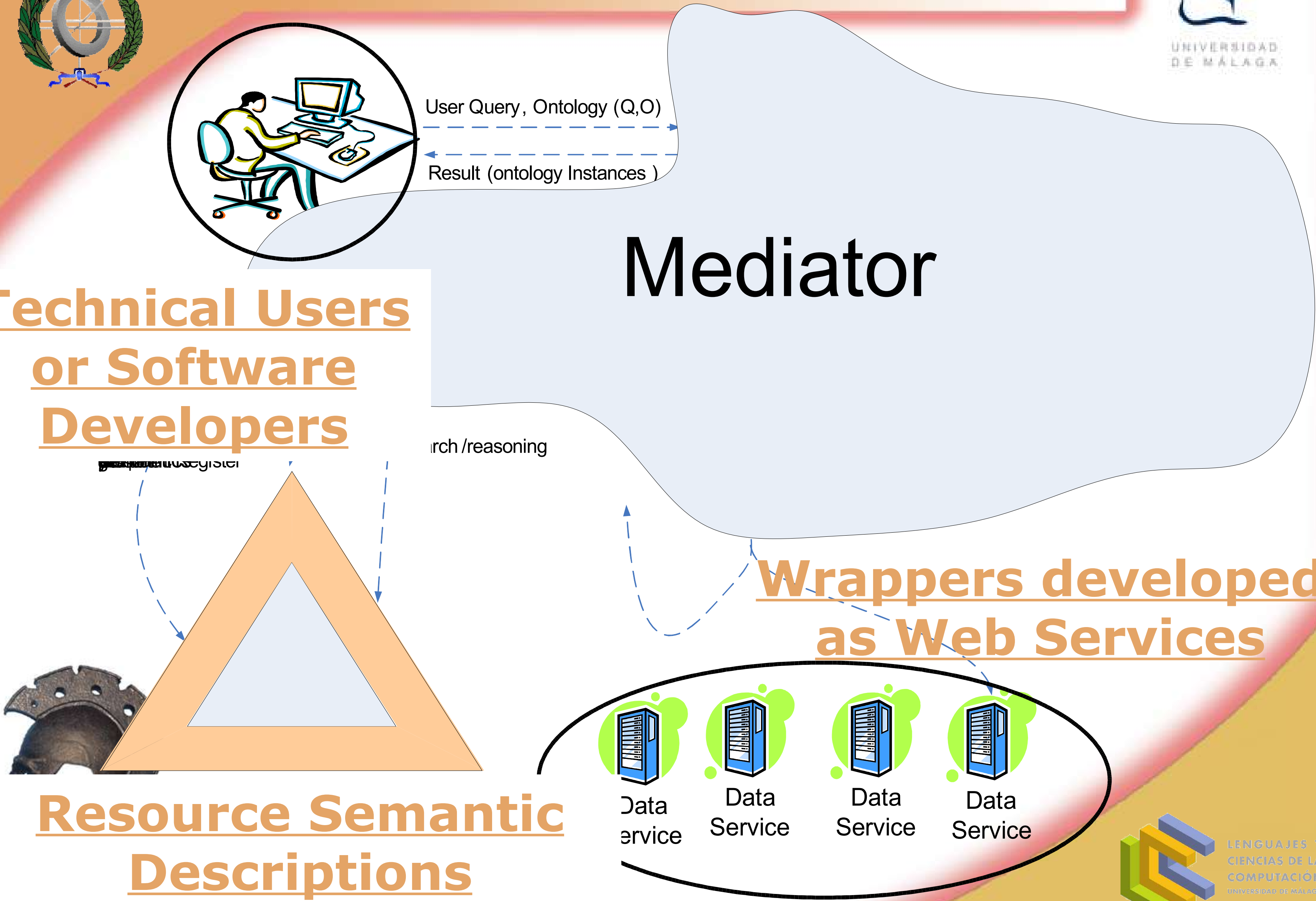
    - Extension → Global Schema Changes

- LAV

    + Easy addition of new data sources

    - Complex Rewriting Process → simpler components will allow partial improvements

Main Characteristics

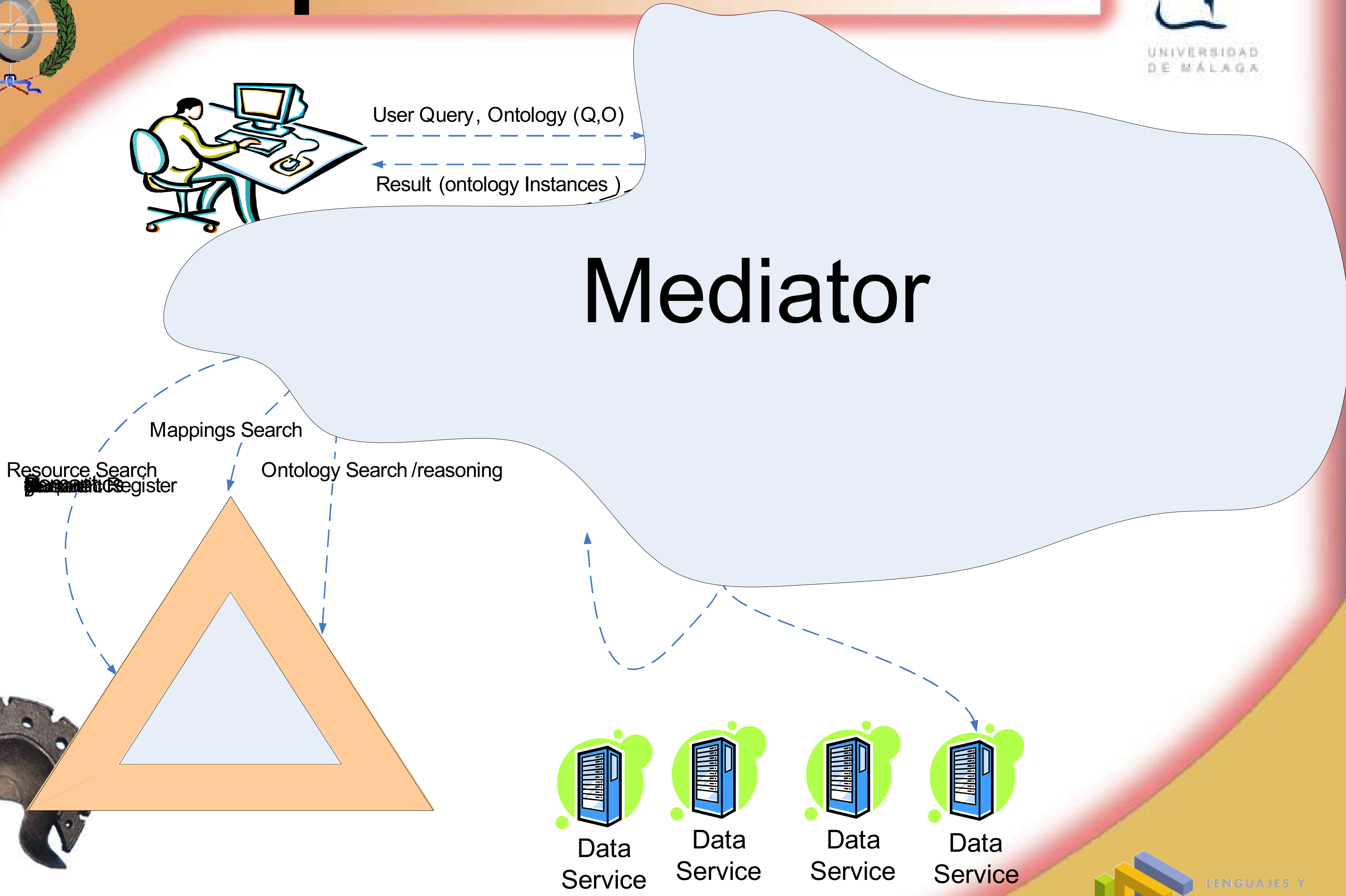


Technical Users or Software Developers

Wrappers developed as Web Services

Resource Semantic Descriptions

Component Division

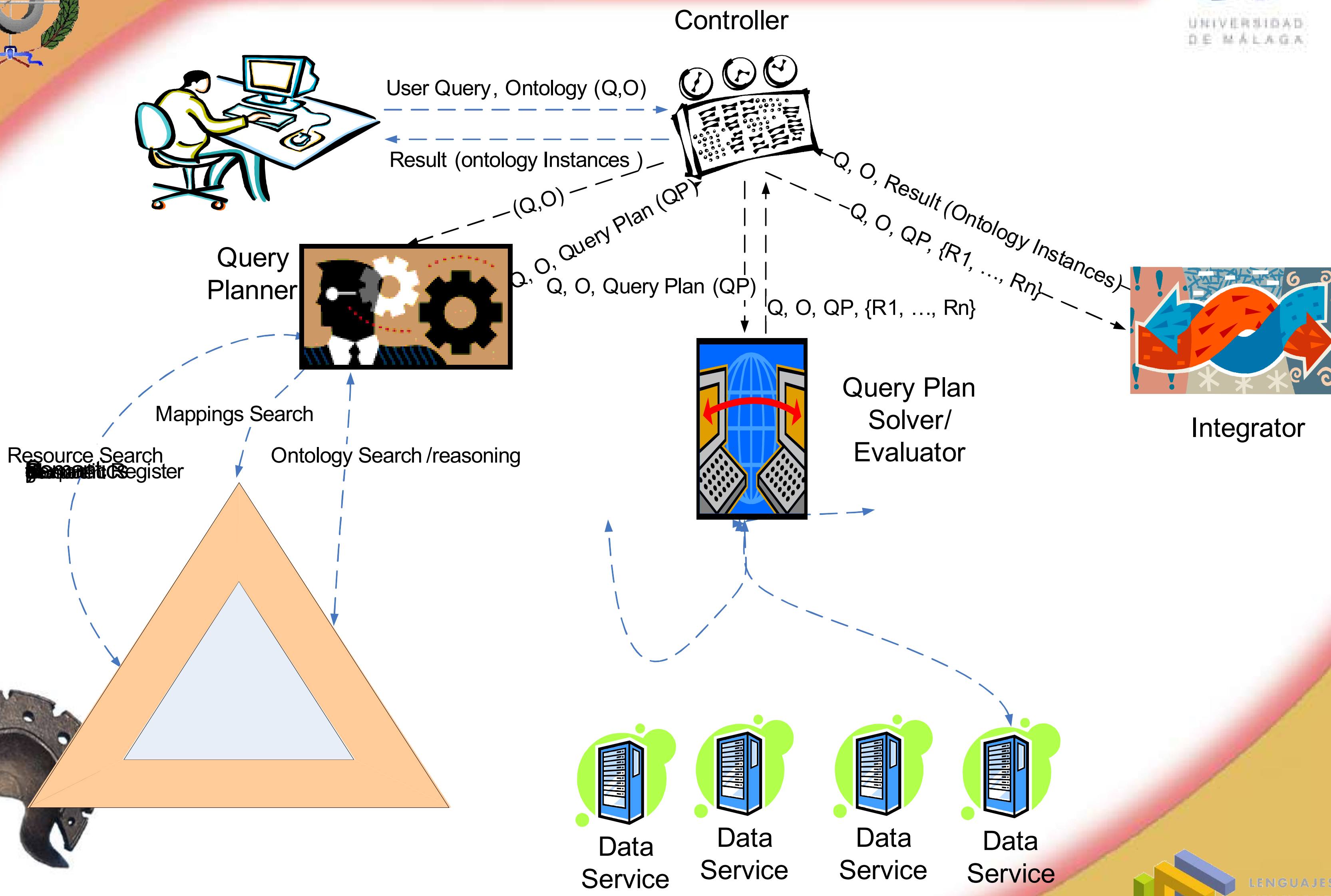


UNIVERSIDAD DE MÁLAGA

K-h-a-o-s INVESTIGACION

LENGUAJES Y CIENCIAS DE LA COMPUTACIÓN UNIVERSIDAD DE MÁLAGA

Partial Improvements



Mediator

interact with the user interface

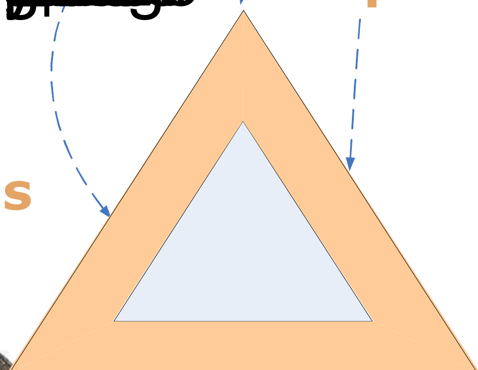
LAV

Ontology Instances

find a query plan (Q for th user query

performs the corresponding call to the data services involved in the sub-queries

Results from data services (R1, …, Rn) are composed



Mappings

Reuse Elements

Share Semantics

Reuse sources

XQuery and XML

Web Services

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

Outline

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

K·h·a·o·s INVESTIGACION

UNIVERSIDAD DE MÁLAGA

LENGUAJES Y CIENCIAS DE LA COMPUTACIÓN UNIVERSIDAD DE MÁLAGA

Use Case

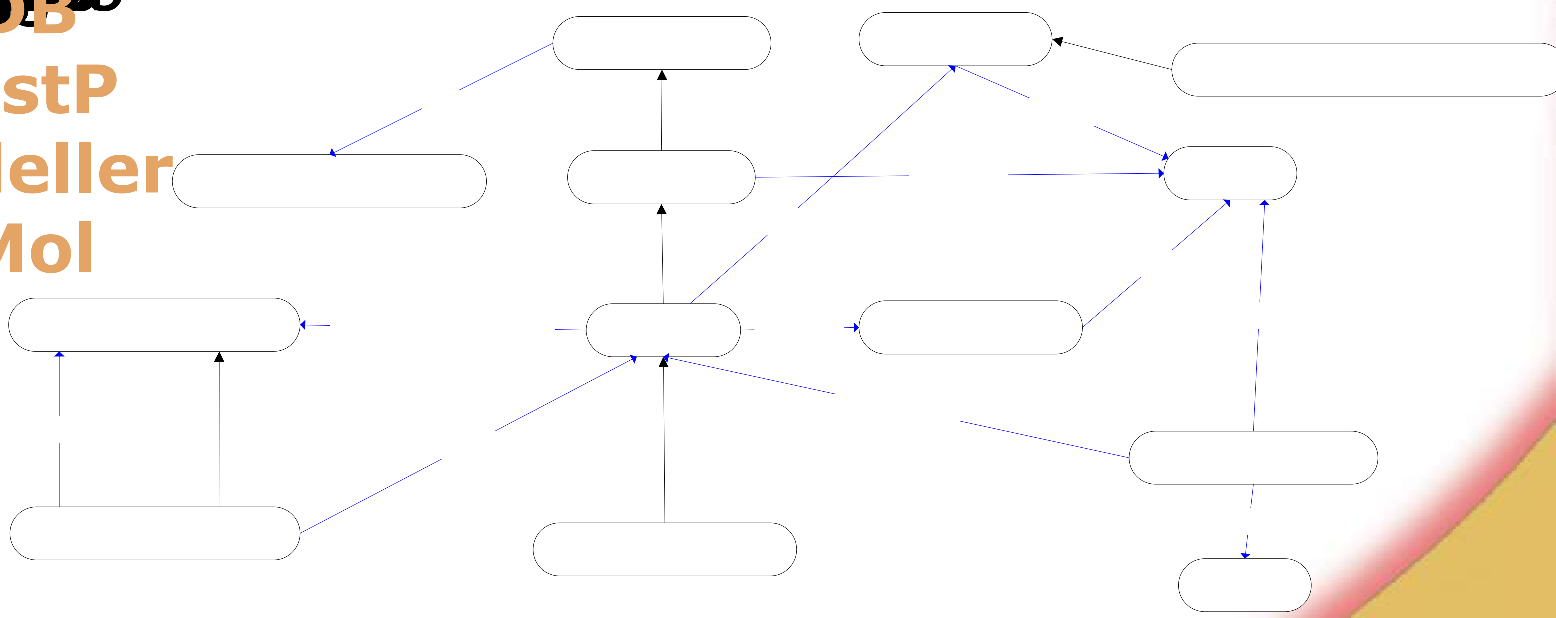
ASP:
Amine System Project
http://asp.uma.es

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

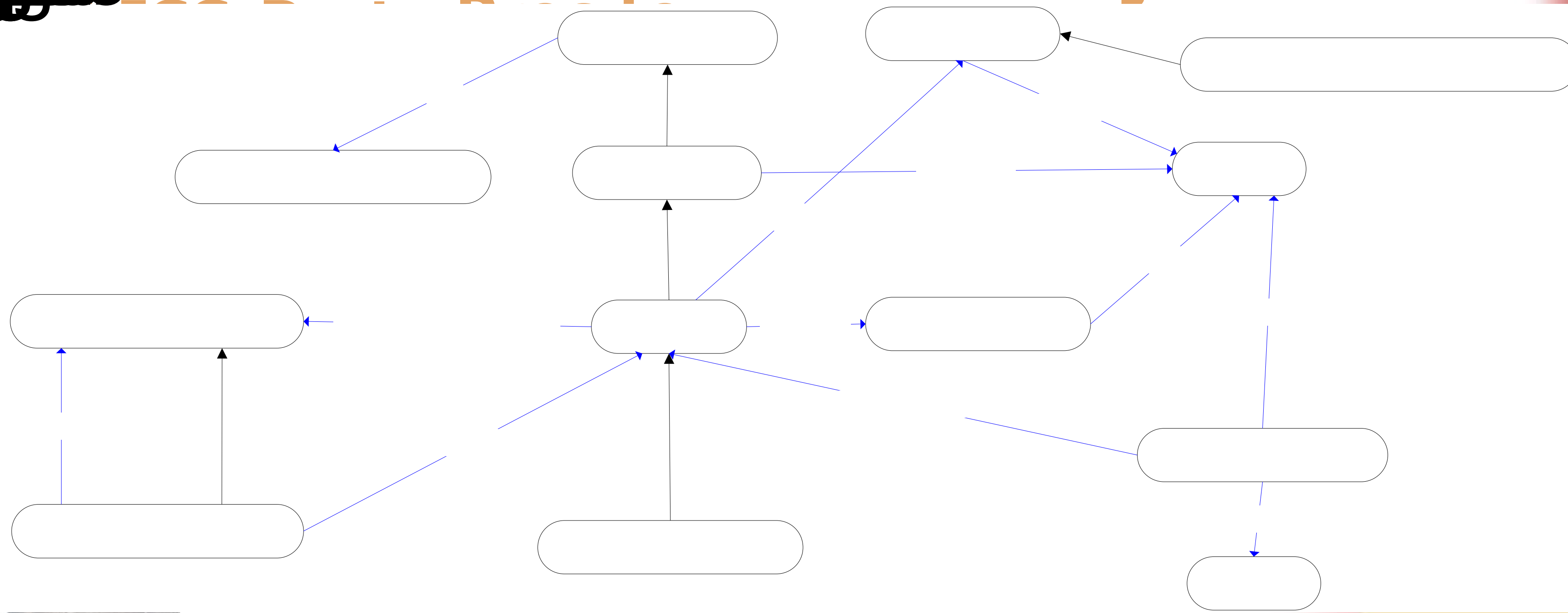
SWISS-Prot
PDB
BlastP
Modeller
JMol



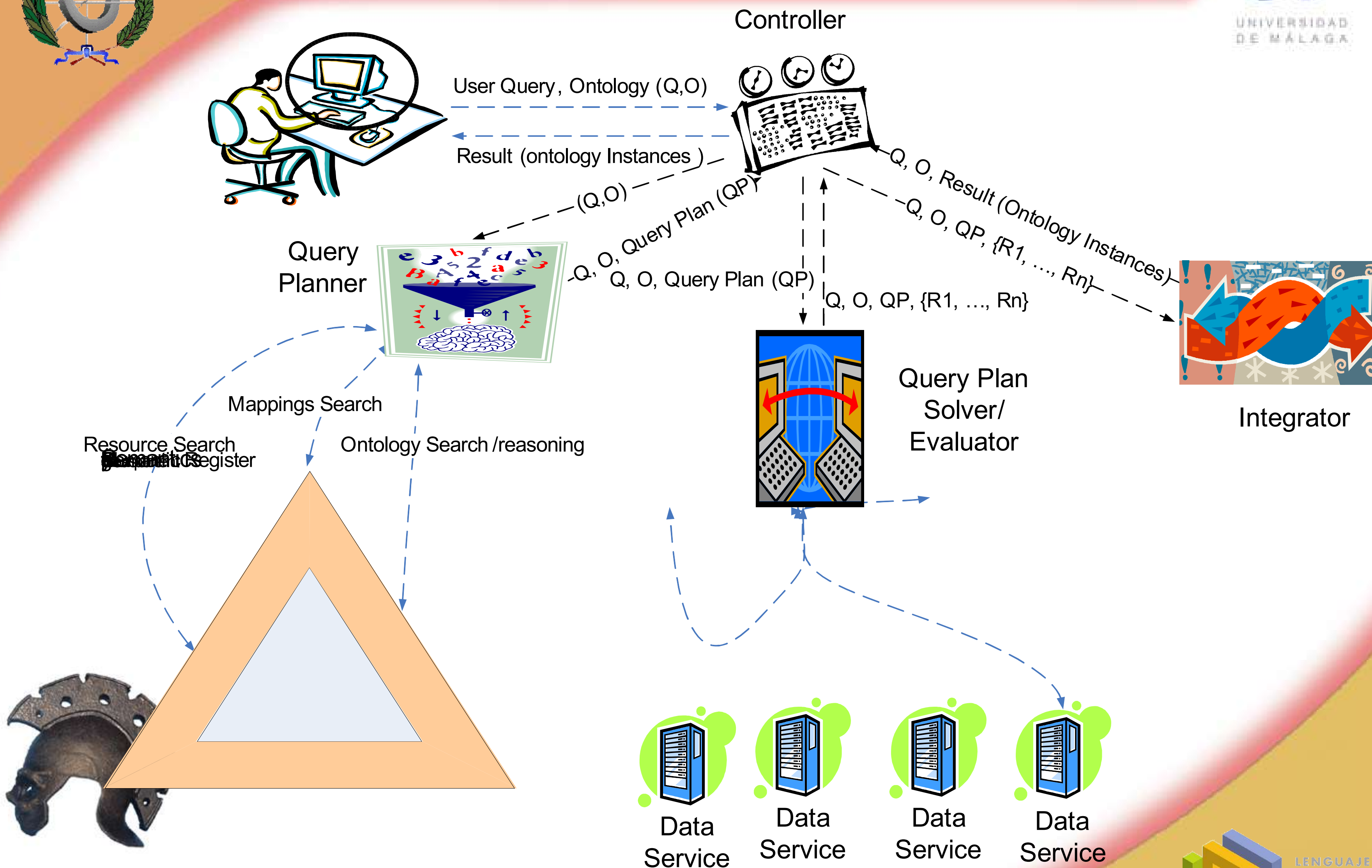
Domain Ontology

PubChem
Kegg
Brenda

ASP:
Amine System
Project

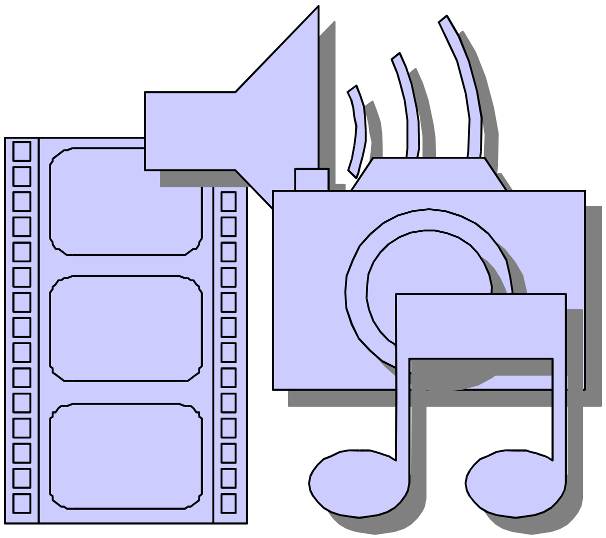


Next Step



Controller
User Query, Ontology (Q,O)
Result (ontology Instances)
(Q,O)
Q, O, Query Plan (QP)
Q, O, Query Plan (QP)
Q, O, Result (Ontology Instances)
Q, O, QP, {R1, …, Rn}
Q, O, QP, {R1, …, Rn}
Query Planner
Query Plan Solver/ Evaluator
Integrator
Mappings Search
Resource Search
Ontology Search /reasoning
Data Service
Data Service
Data Service
Data Service

Application

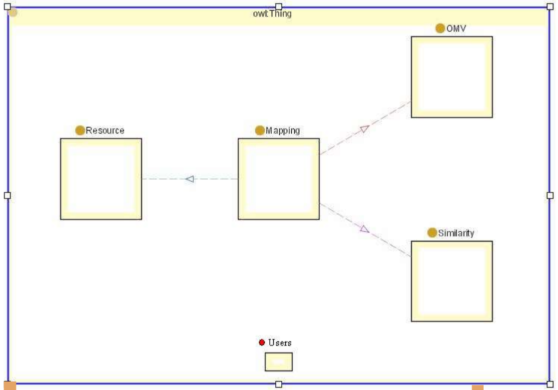


Outline

- 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

<u>Use Case</u>: 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 on-going 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

THANK YOU !.

K·h·a·o·s
INVESTIGACION

UNIVERSIDAD DE MÁLAGA

LENGUAJES Y CIENCIAS DE LA COMPUTACIÓN
UNIVERSIDAD DE MÁLAGA