Database Interoperability through Web Services and Ontologies

Timos Sellis
Institute for the Management of Information Systems
timos@imis.athena-innovation.gr

Dimitrios Skoutas
National Technical University of Athens
dskoutas@dlab.ece.ntua.gr

Konstantinos Staikos
European Dynamics SA
kstai@eurodyn.com
Outline

- Challenges
- Our view for interoperability
- Case study
Today’s landscape

- Biological data distributed in different, heterogeneous databases
- Search and integration is done manually, writing ad hoc scripts
Our perspective on interoperability

Syntactic & Semantic Interoperability

RDF / RDF-S / OWL (semantics)
Our perspective on interoperability

- Data sources are exposed through Web Service interfaces
- Services are discovered and composed
- Information is integrated through ETL processes
- Ontologies provide semantics end-to-end
Web services

- Pieces of software accessible over the Web
- Interoperability over heterogeneous platforms
- Encapsulate data
  - hide implementation details
  - loose coupling
  - facilitate maintenance and evolution (related to paper by Chandras)
- Standard languages and protocols for exchanging data
  - XML, SOAP, HTTP
- Standard languages and protocols for description and discovery
  - WSDL, UDDI
Web services in bioinformatics

Potential benefits

- use as interfaces to biological databases
- unified way to access multiple, distributed biological databases
- provide registries for publishing and searching new data repositories and services
- integrate applications for processing biological data
Web services in bioinformatics

- Survey results from *paper by Hancock*
  - less than 50% of the EC-funded databases currently had Web services available
  - ~30% declared intention to implement Web services
  - 25% - 50% declared intention not to!

- Reasons?
  - lack of awareness?
  - lack of technical knowledge and tools?
Semantic Web services

- Semantic Web services will automate
  - discovery
  - composition
  - invocation

- Our work focuses on effective and efficient techniques for the discovery and selection of Semantic Web services
Semantic interoperability

- Semantic heterogeneity
  - naming conflicts
  - scaling conflicts
  - confounding conflicts

- Using ontologies for semantic interoperability
  - single (global) ontology
  - multiple (local) ontologies
  - hybrid approaches
Extract-Transform-Load Processes

- Traditionally used in data warehouses
Extract-Transform-Load Processes

- The same basic idea underlies other application types, e.g. mahsups (*mashups in bioinformatics?*)
Extract-Transform-Load Processes

- Potential benefits
  - transform data between different formats
  - filter data
  - clean data
  - aggregate data

- Our work focuses on the use of ontologies to facilitate the conceptual design of ETL processes
  - uttermost goal: design ETL processes declaratively
Case study (K. Staikos – MSc Thesis)

- Goal: develop Web services to integrate data from 3 biological databases:
  - EMBL
    - nucleic acid sequence data
  - MEDLINE
    - bibliographic database for medicine, nursing, health services, etc.
  - ArrayExpress
    - annotated microarray data
Case study

- A Web application was developed using Web services to integrate data from the aforementioned 3 databases to a target repository
  - technologies used: XML, SOAP, EJB, Axis, Servlets/JSPs, Struts
- The project (MSc thesis, TU Munchen) was implemented in 2004
Conclusions

Mature technologies now exist to face the problem of interoperability

Tighter collaboration between bioinformatics people and database people will benefit both communities