IBM Centers for Advanced Studies

Centro Studi Avanzati di Roma

Semantics in data integration processes

Guido Vetere, Research Director
IBM Center for Advanced Studies of Rome

NETTAB 2005
Napoli, October 4-7 2005

Summary and motivation

- **A huge amount of biological information and many bioinformatic systems are available over the Internet**
- **Making both data and systems interoperable in coordinated workflows would unleash new scientific capabilities**
- **Biological information is dealt by means of “heterogeneous data structures and information systems and, often, even a different semantics” (Paolo Romano)**
- **Web Services are able to neutralize differences in platforms and data structures, but are noncommittal about semantics**
- **This lecture is about WS semantic integration, with some specific reference to Bioinformatics**

Interoperability

- **Interoperability is when autonomous systems can exchange data and activate functionalities transparently, reliably, and securely across a network**
- **Web Services standards provide today a solid and widely accepted platform for system interoperability**
- **Complex Life Science computations, data services, and workflows could better leverage distributed architectures, if basic functionalities were exposed through Web Services**

Examples: Life Sciences WS at IBM's alphaWorks

- **GenBank:** *queries a Web database at the National Center for Biotechnology Information (NCBI) Web site and returns the nucleotide sequence for each accession number submitted.*
- **BLAST (Basic Local Alignment Search Tool):** *conducts a sequence alignment analysis for each input sequence at NCBI*
- **ClustalW:** *runs a fully automatic program at the European Bioinformatics Institute (EBI) for global multiple alignment of DNA and protein sequences*

From Interoperability to Integration

- **Integration is when a set of interoperable systems are coordinated to act as they were a single one**
- **Moving from Interoperability to Integration requires harmonizing data and processes semantics**
- **However, in their basic form, Web Services are neutral w.r.t. semantics**
- **Industry and research are striving to provide WS infrastructures (and the Web in general) with a (sort of) 'semantic layer'**

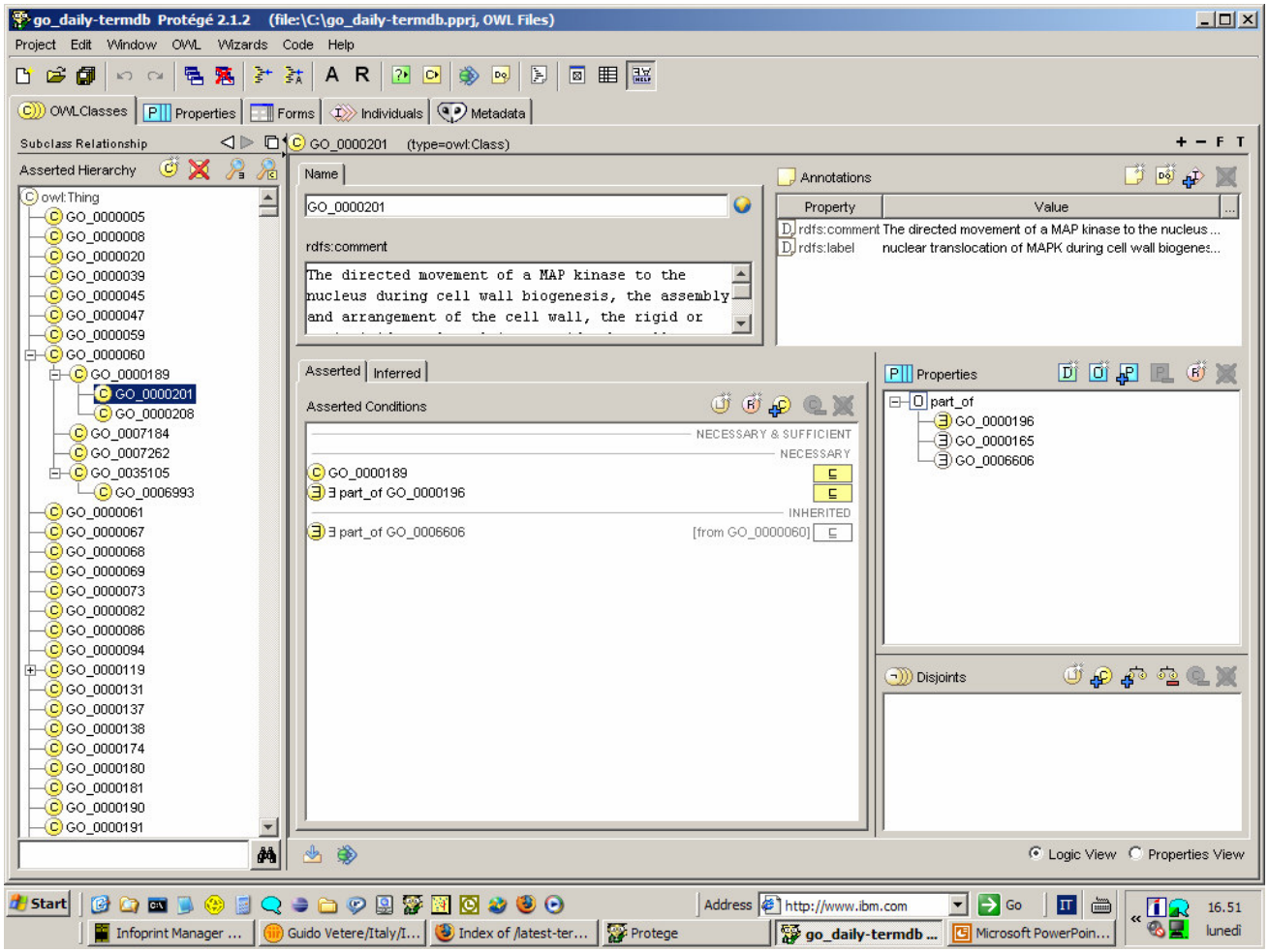
Semantics

- **In general, semantics is a mapping (aka 'interpretation function') which involves:**
  - Expressions: a system of manifested symbols (e.g. a formal language)
  - Contents: a system of something else which is not necessarily apparent (e.g. sets of objects or events in (some abstraction of) the 'Real World')
- **Web Services 'semantics' aims at filling the gap between:**
  - Expressions: the description of operations and data items (WSDL,XML)
  - Contents: the (interpretation of) some shared conceptualization
- **A number of WSDL extensions (e.g. WSDL-S), along with rich XML-based schema modeling languages (e.g. XSD, XMI, RDFS, OWL) are available to implement WS semantic extensions**

Semantic issues

- **Unfortunately, in its generality, semantic integration is essentially a non-technical issue: the notion of 'sharing a conceptualization' (i.e. an 'ontology') involves deep and controversial philosophical aspects**
- **Working out broad ontologies requires extensive and complex analyses and many discretional and debatable choices**
- **Adopting available ontologies involves (costly) social & technical adaptations**
- **Fortunately, Life Science WS can benefit of a vast array of ontologies (e.g. Open Biomedical Ontologies, Gene, etc) mostly based on well-understood *natural kinds* and *processes***

Example: Gene Ontology (OWL, Protégé)



Semantics in services infrastructures

- **Since the adoption of shared ontologies is a long and complex process, biologists will probably have to handle "heterogeneous data structures and […] different semantics" for many years**
- **This requires semantic integration to resort on *conceptual mappings* that make different data/process descriptions equivalent, either pair-wise or with respect to some (partial) unifying ontology**

Conceptual mappings

- **Abstractly, a conceptual mapping is a formula**

$$\forall \mathbf{x}\ \exists \mathbf{y}\ \Phi\ (\mathbf{x},\mathbf{y}) \rightsquigarrow \exists \mathbf{z}\ \ \Psi\ (\mathbf{x},\mathbf{z})$$

  - **x**, **y**, and **z** are variable vectors
  - Φ and Ψ are conjunctive formulas of predicate symbols belonging to different alphabets (ontologies)
  - $\rightsquigarrow$ is a logic implication connective (e.g. FOL material implication)
- **Concretely, mappings can be:**
  - any kind of XML transformation rule (e.g. XSLT)
  - specific assertions of ontology languages (e.g. OWL's *sameClassOf*)
  - Named views in database federations
  - …

Four models for semantic integration

- **Depending on whether**
  - conceptual mappings are drawn toward a single unifying ontology (model) or not
  - their evaluation is distributed or centralized

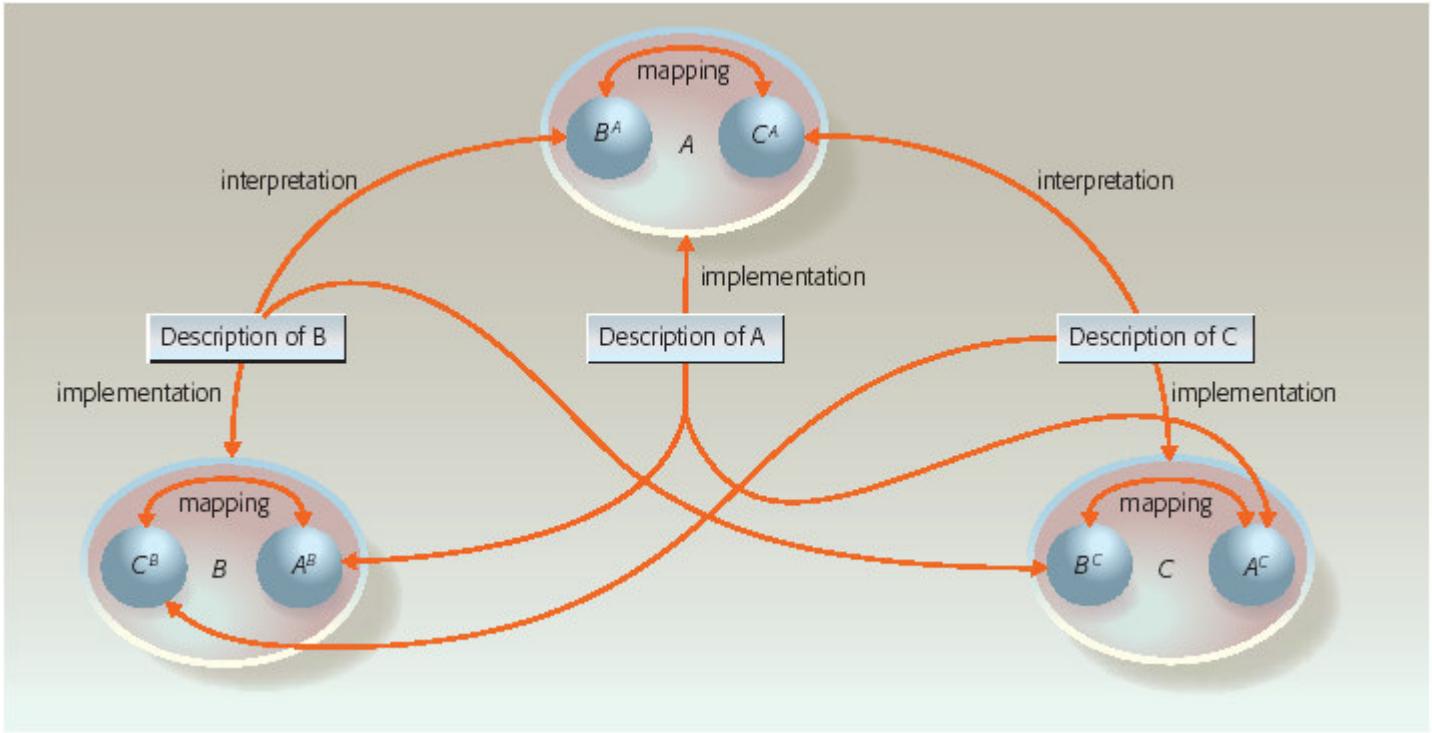
  we have the following four basic models for semantic interoperability:
  - unmodeled-decentralized
  - unmodeled-centralized
  - modeled-decentralized
  - modeled-centralized

(Read the full story in: G.Vetere, M.Lenzerini, Models for Semantic Interoperability in Service Oriented Architectures, IBM Systems Journal 44, Oct. 2005)

Unmodeled-decentralized

The integration logic is distributed, and there are not shared ontologies

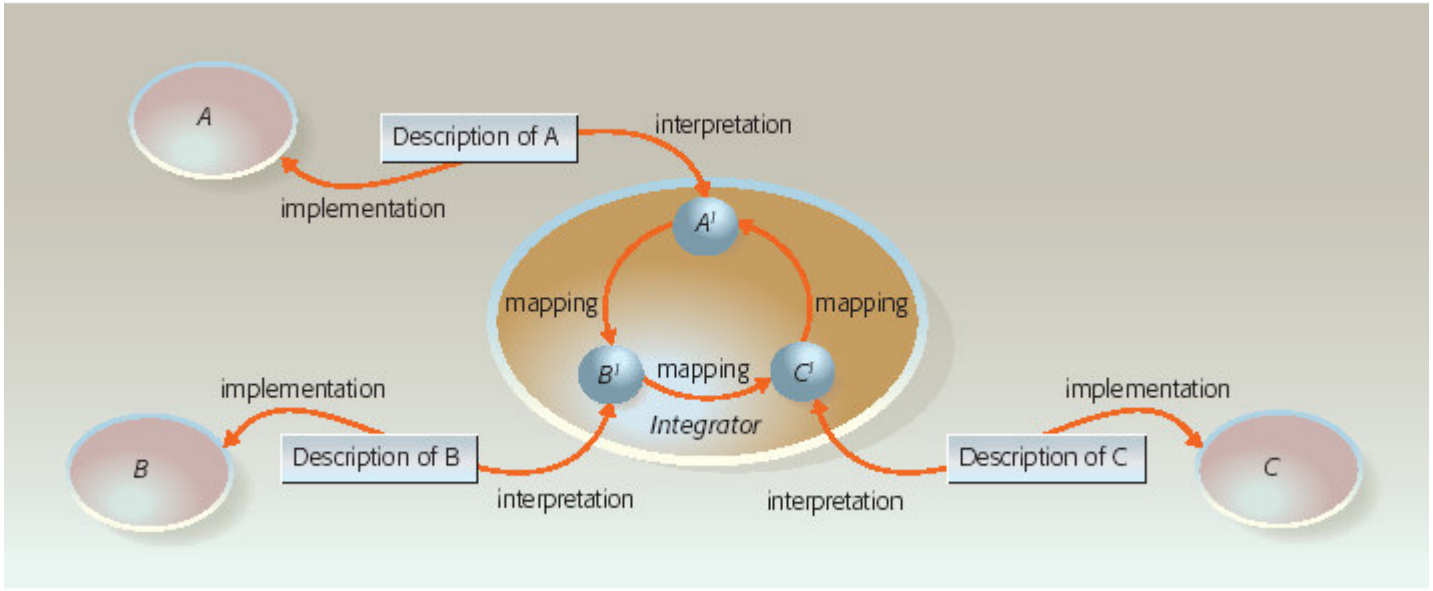
- ‘Pure’ Peer-to-Peer systems, the Web ‘as is’,
- P2P information integration systems, ‘emergent semantics’



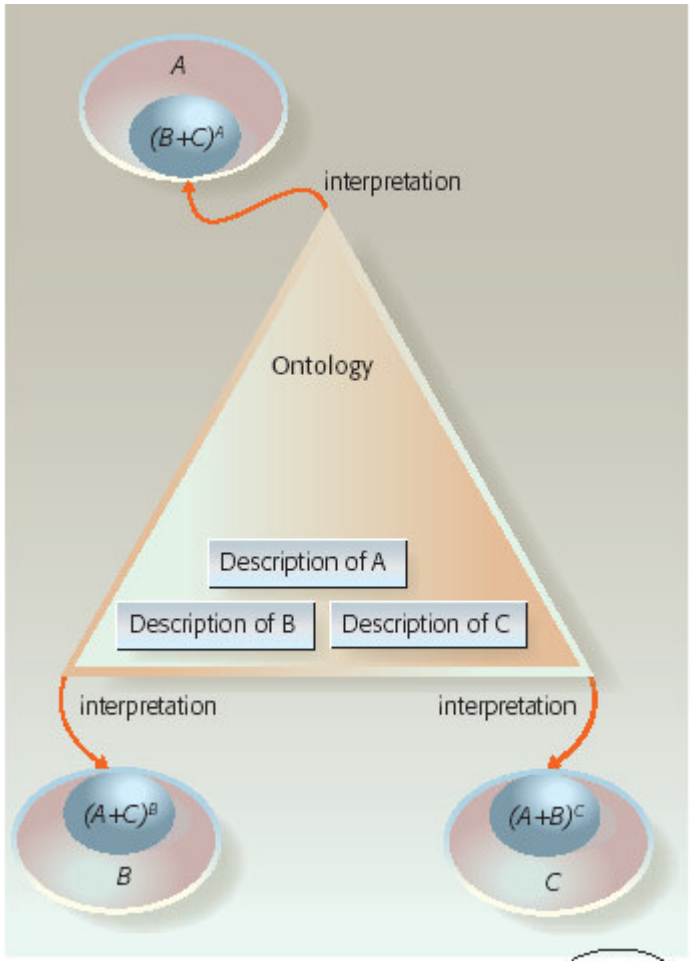
Unmodeled-centralized

The integration logic is contained in a single system, without an explicit unifying ontology

- Choreographies, *ad hoc* Data Grid applications ('analysts'), …
- BPEL, OGSA-DAI, …



Modeled-decentralized semantic integration



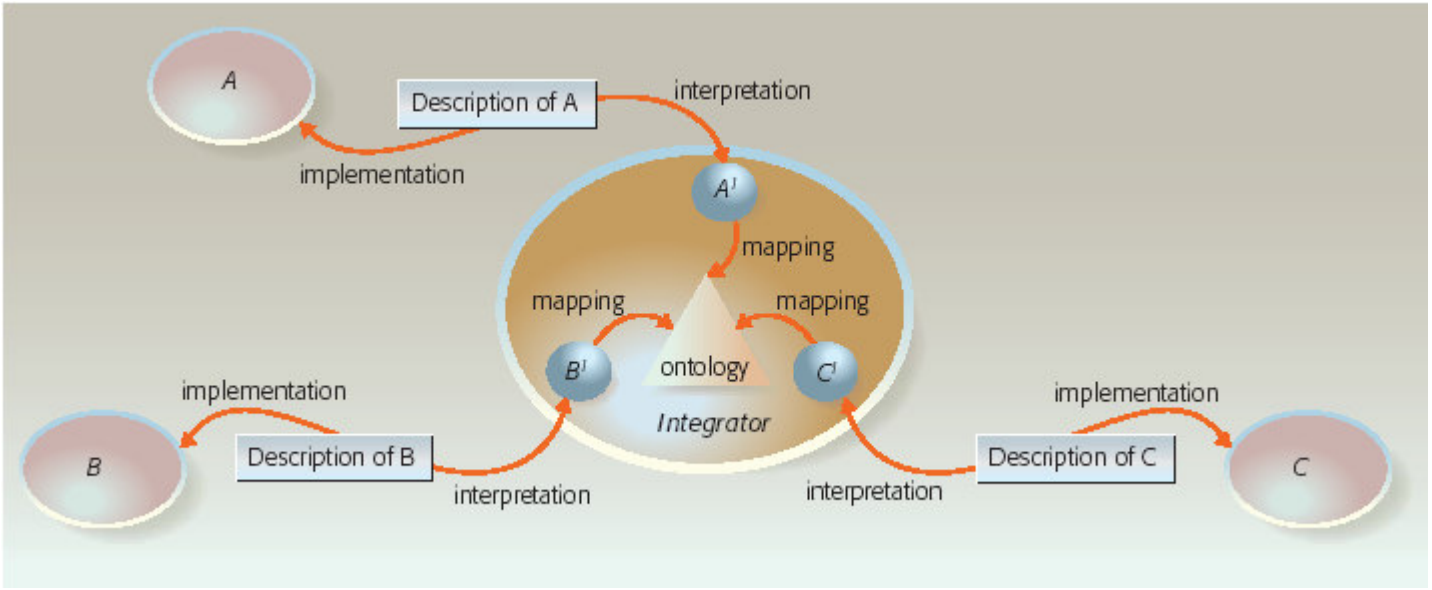
The integration logic is distributed by any service implementation, based on a shared ontology

- 'Semantic Web' approach
- Model-Driven Web Services *à la* WS Modeling Framework, Semantic Overlay Networks, Semantic Grid, etc

Modeled-centralized semantic integration

The integration logic is contained in a single system, based on a unifying ontology

- Classic database federation and ‘semantic’ data grids
- Industry application integration infrastructures based on ‘business models’



Discussion

- **Despite many open 'philosophical' issues, ontologies allow (but not ensure) semantic integration, whatever it could be in practice**
- **However, there are many cases in which organizational, cultural, or infrastructural constraints hinder or even disallow the adoption of such semantic artifacts**
- **In Bioinformatics, the availability of stable natural taxonomies, definitions, theories, etc is certainly an excellent starting point for modeling ontologies – and in fact there are many**
- **In their maturity, these standards could enable semantic integration through Web Services in both centralized and decentralized infrastructures**

Conclusion

- **The availability of mature semantic standards emerges as an important requisite for the development of distributed workflows in Bioinformatics**
- **Creating Bio-ontologies and/or experimenting them in concrete data and process integration is a (non-trivial) work in progress, that involves both theoretical and practical aspects**
- **Assessing / improving the quality of Bio-ontologies is a priority**
- **Application to Bio-ontologies of specific development methodologies (e.g. Guarino&Welty's *OntoClean*) and basic ontological distinctions (e.g. CNR's DOLCE) deserves a deep investigation in the coming years**

Questions?

CAS Centers: http://www.ibm.com/ibm/cas

Annual conference: http://www.ibm.com/ibm/cas/cascon