Semantics in data integration processes

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

Integrating information services in an ever growing internetworking World is likely to be the most urgent need for any kind of business, trade, or science. Research in biology is not an exception: a huge and increasing amount of complex and heterogeneous biological data is distributed over a network of computing system, which could interoperate in coordinated workflows to support complex and extensive investigations. Thus, data and process integration emerges as an interesting topic in bioinformatics: the goal is to unleash the potential of system interoperability and integration in the specific field of biological research.

Computer Science has been focusing on interoperability and integration for decades. From the perspective of interoperability, Web Services standards and technologies have reached today a reasonable level of maturity, and their acceptance is constantly growing. Thanks to their binding to W3C standards, Web Services provide a commonly accepted platform that allows systems to exchange data and activate remote functionalities transparently, reliably, and securely across the Internet. For the community of researchers in biology, this amounts to the possibility to share data and computational resources, thus getting the ability to perform computational tasks in a cooperative manner and, ultimately, to reach new research frontiers.

Moving from Web Services interoperability to the concrete integration of processes and data, however, is neither immediate nor easy. In fact, Web Services provide support at the level of protocols, syntaxes, and infrastructural services, but are neutral with respect to data and processes semantics. To give an idea, using Natural Language as a model, Web Services provide grammar rules, while most of the linguistic competence required for communication is concerned with lexicon, that is, in the end, with semantics. In this light, it is not surprising that the need of supplementing a 'semantic layer' to Web Services infrastructures (and to the Web in general) is so strongly and diffusely felt [1]. As a matter of facts, since 1998, when Berners-Lee exposed the vision of a 'Semantic Web', an uncountable number of research, standardization, and industrial initiatives have been promoted with the aim of achieving a suitable treatment of semantics for Internet-based information gathering and exchanges, including those in the Life Sciences [2].

Roughly speaking, semantics is a mapping between a system of symbols (e.g. a formal language) and a system of anything else (e.g. sets of objects and events). For Web Services, semantic specifications aim at filling the gap between the description of operations and data items on one hand, and the portion of reality in which the Service is grounded on the other hand. Note that, in very basic scenarios, this gap could be naively filled based on the "natural language flavor" of symbols used for the service's description (e.g. WSDL operation names). However, it is generally recognized that, in order to effectively exploit Web Services in real business settings where many independent actors are involved, semantic specifications must be explicit, accurate and consistently shared between parties. This can be achieved by modeling anything the service refers to by means of special kinds of logic theories called 'ontologies'. Ontologies, in brief, are formal accounts of "what exists" in a certain domain, commonly based on set-theoretical concepts such as classes (e.g. *amino-acid*), relations (e.g. *part-of*), and individuals (e.g. *this-substance*). Specific languages, such as OWL, have been designed to convey ontologies and share them over the Web by means of W3C standards, and a number of enabling technologies are available to support them [3].

Despite the increasing availability of semantic-oriented standards and technologies, the problem of dealing with semantics in Web-based cooperation, taken in its generality, is very far from trivial, not only for practical reasons, but also because it involves deep and controversial philosophical aspects. Nevertheless, for relatively small communities dealing with well-founded disciplines such as biology, concrete solutions can be effectively put in place. In fact, most of the data structures will represent commonly understood *natural kinds* (e.g. microorganisms), well-studied *processes* (e.g. syntheses) and so on. Still, significant differences in the way actual data structures are used to represent these concepts might require complex mappings and transformations. In the sequel, we will survey the basic models for semantic interoperability in service-based infrastructures with some remark about the specific context of bioinformatics.

A service-based infrastructure is basically a set of systems acting either as consumers or providers, or both. Providers hold data and support standardized access and manipulation operations on them. Services are described in terms of data and operational schemas. These descriptions, along with other information, are stored into registries where consumers can retrieve them. Then, consumers can connect to providers and finally perform operations such as queries or updates, through suitable invocations. In these infrastructures, semantic interoperability means that the set of descriptions stored into registries are interpreted in a consistent way across the entire infrastructure. Since providers and consumers are, in principle, independent organizations, and the descriptions they deal with are not necessarily bound to the same ontology, semantic interoperability requires a kind of integration that ultimately consists in a *conceptual mapping* that makes different descriptions equivalent, either pair-wise or with respect to some unifying ontology.

Abstractly, conceptual mappings can be expressed as correspondence rules of the form:

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

where **x**, **y**, and **z** are variable vectors, Φ and Ψ are conjunctive formulas of predicate symbols belonging to different descriptions, and → is a logic implication connective (e.g. first-order material implication). Concretely, mappings are any kind of transformation rules such as XSL transformations, specific assertions made by means of ontology languages (e.g. OWL's *sameClassOf*), or configuration data for query reformulation technologies. Now, depending on whether conceptual mappings are drawn toward a single unifying ontology (let's call it *model*) or not, and whether their execution is distributed or centralized, we have the following four basic models for semantic interoperability [4]:

- modeled-centralized
- modeled-decentralized
- unmodeled-centralized
- unmodeled-decentralized

Modeled-centralized semantic integration is one in which a common ontology is adopted and the integration is performed by a single system. This is the model of classic database federation and data grids, where queries are posed against a specific integration system hosting a 'global virtual view' that reconciles the heterogeneous schemas exported by a set of distributed sources. Also, this is the model adopted by industry-level integration infrastructures based on the notion of 'service bus' [5], where all the messages are managed by a centralized component that is able to transform and route them according to the correspondence of their content with respect to a 'business model'.

Modeled-decentralized semantic integration is also characterized by a common ontology, but, since there is not a single integration system, any service is requested to implement the unified semantics someway. Typically, this is the approach adopted by the 'Semantic Web', where a widespread adoption of well-understood and shared ontologies is envisioned. Web Services Modeling

Framework [6], for instance, adopts this model to allow (but not ensure) different services to implement an homogeneous semantics.

Unmodeled-centralized semantic integration is one in which the integration is achieved in a single system without an explicit semantic model. Although it might seem an oddity, this is actually the basic model for Web Services choreography systems. In fact, Web Services choreographies based on specific languages such as BPEL consist in procedures (workflows) in which a set of heterogeneous services are invoked in a coordinated way. In these applications, semantics consists in the way the output of services' invocations is used as an input for other services and to drive the overall system behavior. Up to now, the way this is accomplished is left to the implementation, and there are no requirements regarding the adoption of any specific artifact to contain the integration semantics.

Unmodeled-decentralized semantic integration takes place when each system adopts its own ontology, is responsible for establishing its own conceptual mappings with anyone else, and performs its own integration logics. Not surprisingly, it is very difficult to get to semantic integration in this situation, but, as matter of facts, this is the normal condition in really 'loosely coupled' environments such as the Web in general. Hence, studies and experimentations have been recently started to understand how to cope with semantics in this sort of peer-to-peer, relativistic setting [7].

It is easy to observe that both centralization and modeling facilitate Web Services developers in setting out a common semantics. However, there are many cases in which organizational, cultural, or infrastructural constraints hinder or even disallow the adoption of such policies. As for bioinformatics concerns, the availability of stable taxonomies and scientific definitions is certainly an excellent starting point for modeling semantic standards such as a coordinate set of shared ontologies. If widely adopted and correctly implemented, these standards would allow distributed and heterogeneous systems to cooperate through Web Services in both centralized and decentralized infrastructures. Nonetheless, the conception of biological ontologies, well suited for Web Service integration, is still underway [8].

In conclusion, the availability of Web-exploitable semantic standards (e.g. a set of OWL ontologies) for biology emerges as a fundamental enabling condition for the development of distributed workflows in bioinformatics. Admittedly, this conclusion is not new: the work around the Gene Ontology [8], for instance, began in 1998. However, when looking at the state-of-the-art, it seems that much has still to be done to make current ontological resources suitable for concrete data exchanges and processes integration based on Web Services. This is a very important research task that the bioinformatics community should pursuit in the next future.

References:

1. Semantic Web site at W3C: http://www.w3.org/2001/sw/
2. Semantic Web for the Life Sciences: http://www.biopathways.org/semweb/
3. I. Herman, W3C, Tutorial on Semantic Web Technologies http://www.w3.org/Consortium/Offices/Presentations/RDFTutorial/
4. G. Vetere and M. Lenzerini: Models for Semantic Interoperability in Service Oriented Architectures, IBM Systems Journal 44, 2005
5. Patterns: Implementing an SOA Using an Enterprise Service Bus, IBM RedBook 2004
6. D. Fensel and C.Bussler, The Web Services Modeling Framework WSMF, http://www.wsmo.org/papers/publications/wsmf.paper.pdf

7. D. Calvanese et al, Logical Foundations of Peer to Peer Data Integration, Proceedings of SIGMOD 2004.
8. The Gene Ontology, http://www.geneontology.org