Injecting Semantics into Scientific Collaboration

Muthukkaruppan Annamalai[1], Lito Cruz[2], Leon Sterling[1], Glenn Moloney[1]

[1]The University of Melbourne, Victoria 3010, Australia
[2] RMIT, Victoria 3000, Australia

INTRODUCTION

The WWW has become the *defacto* collaborating medium for distributed scientific community to interchange information among them. There is still much human mediation involved to utilise this information. The human effort can be largely reduced, when the information is exchanged with meanings attached. The key enabler for this meaningful collaboration on the Semantic Web is ontology.

An ontology does not have to be a universally standardised language. However, its usability depends chiefly on its adoption as a collaborating language by a user community. Following this, our research aims to demonstrate that suitable ontologies can be constructed to support effective web-based collaboration within a distributed scientific community, such as the Experimental High-Energy Physics (EHEP) collaboration.

INFORMATION SHARING WITHIN THE EHEP COLLABORATION

The EHEP collaborative work revolves around experimental analyses. The research groups within a collaboration analyse the huge sets of data produced in an experiment, using various analysis techniques. The results of the analyses are communicated to fellow researchers in the form of pre-prints and research notes.

In the absence of a prescribed set of analysis description guidelines, authors generally state aspects of the analysis procedure, which they think is essential to be conveyed to the readers. As in the case of experimental science publications, there is a tendency among authors to presume readers already have knowledge about the analysis procedure. The Experimental analyses described in this fashion, with publication bogged down with tacit knowledge are prone to be misunderstood, particularly by researchers who are not familiar with the kind of analyses mentioned in the document. Often times, a researcher trying to replicate published experimental analyses, ends up with relatively different result. Precious time is expended trying to correctly interpret the experimental analyses, which often results in tedious debugging of the analysis procedure. Debugging an experimental analysis described by authors who profess somewhat different ontological commitment about the domain is indeed a daunting task.

EXPLICATING THE EHEP ANALYSES

The problem can be traced to lack of structure and semantics in the published scientific documents. We believe it can be safely resolved if an analysis process is described explicitly in definite terms to peer researchers. To begin, we propose the creation of a formal scientific document, called *analyses report*, which describes the completed experimental analyses according to EHEP ontologies in an orderly manner. An organised narrative of the analyses would allow a meaningful description of the content. Publishing the analyses with annotations linked to ontologies published on the Semantic Web can ensure optimal exchange of information between researchers within a collaboration.

The EHEP ontologies can be used to mark-up the essential parts of the publications in open archives, allowing semantic searches on the collection. Alternately, a publication can now straightaway point to the relevant experimental analysis reports in the analysis archive. Accessing relevant publications or discovering similar experimental analyses will require far less time and effort.

Moreover, these machine-readable ontologies can also be utilised to describe analysis jobs, which can be directly processed by *analyser* agents to perform the required analyses. This opportunity to embark upon an innovative way of handling scientific information generated within an EHEP collaboration is illustrated in **Fig. 1**.

AGENTS AS EXPERIMENTAL DATA ANALYSERS

Analysing the EHEP experimental data involves various statistical physics calculations and simulations. At present, the physicists spend much time with coding or modifying their analysis programs. The EHEP physicists would rather focus their attention on analysing physics phenomena than dealing with programming minutiae.

We envision an agent system that can liberate the physicists from drudgery of programming. Currently, there are distributed analysis programs developed by the EHEP physicists, written in FORTRAN and C++. Ideally, these programs should be shared among fellow researchers in a collaboration without much difficulty. Based on the Multi-Agent System model found in the RETSINA project, these routines could be wrapped inside a service provider agent and mapped onto an agent service that can be shared, offered and consumed by other agents on the collaboration's network. As a result, part of the computational analyses done by the physicists can be delegated to agents.

Ontologies are required to provide the source of vocabulary for agents to communicate with one another. In our case, there is a need for two types of ontologies, the domain and service ontologies. The domain ontology captures the intrinsic structure of the EHEP domain embodied as concepts, relations and axioms. Agents use it to interpret the analysis task and also to review the contents of marked-up analysis documents (as described earlier). The service ontology defines the terms pertaining to analysis services. The broker agent and the service provider agents use it as a vehicle for mutually discovering services dynamically. This idea is presented in **Fig. 2**.

BUILDING THE DOMAIN ONTOLOGY

We set out to accomplish this task by first creating a skeletal EHEP knowledge model, such as the one shown in **Fig. 3**. It will serve as the foundation for the domain ontologies. This informal model is elaborated from interviews with EHEP researchers, scientific documents, such as pre-prints and journal articles, and existing standard HEP terminology, such as the terms maintained by the Particle Data Group.

The revision of the 'obscure' informal model into reliable ones, is a gradual process, and is achieved mainly through the meetings with the physicists. Physicists are conversant with their scientific terminology but are not adept with organising ontologies or agents; while computer scientists with experience in these areas are beginners when it comes to the specialised EHEP domain. The interaction reveals the need to help each other understand each other's concerns in the process of constructing this knowledge model, which also encompasses the model validation, verification and refinement.

Finally, the completed model will be formalised as EHEP ontologies. We intend to implement the ontologies in DAML+OIL, which is set to be the standard semantic mark-up language for web resources.

BUILDING THE SERVICE ONTOLOGY

The preliminary task in the construction of the service ontology is distinguishing the candidates for services. We need to identify and classify the phases of work an EHEP physicist goes through and refer them as potential candidates for services. They are then analysed for sub-components, which may be identified as additional services. In the process of identifying services in EHEP analyses, the computer scientists are in effect guiding physicists to think of their analysis steps as composition of service blocks. We have identified some basic analyses which the EHEP physicists perform and within them we have identified candidate services. In **Fig. 4** we provide a snapshot of a common analysis carried out for searching a target particle. The processes designated by a rectangle are candidates for becoming service. Once the services have been identified with good precision, they need to be described. For this, we intend to use the DAML-S service ontology description language.