A Collaborative Framework for Distributed Scientific Groups

Muthukkaruppan Annamalai
Dept. of Computer Science and
Software Engineering,
The University of Melbourne,
Victoria 3010, Australia.
+61 3 83449100
mkppan@cs.mu.oz.au

Leon Sterling
Dept. of Computer Science and
Software Engineering,
The University of Melbourne,
Victoria 3010, Australia.
+61 3 83449100
leon@cs.mu.oz.au

Glenn Moloney
School of Physics,
The University of Melbourne,
Victoria 3010, Australia.
+61 3 83445455
glenn@physics.unimelb.edu.au

ABSTRACT

The potential of collaborative work can be further harnessed if the implicit knowledge in the collaboration documents can be exploited. Together with the Experimental High-Energy Physics (EHEP) community, we are investigating the use of ontologies for scientific collaboration. The EHEP collaborative work revolves around experimental analyses. We propose an intuitive way to establish augmented collaborative experimental analysis documents. The collaboration documents are annotated with appropriate semantic descriptors, linked to ontologies published on the web. Our initiative will necessarily lead the EHEP community to produce and share information innovatively. This development is an epitome of large-scale scientific collaborations and will provide the impetus for a more rapid scientific advance.

1. 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 is ontologies.

Ontologies are a specification of conceptualisation [3] and are in essence, a set of formally defined vocabulary in a shared domain. 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 the exchange of meaningful information within a distributed EHEP collaboration.

2. THE EHEP COLLABORATIVE WORK

EHEP is dominated by large collaborations with membership from all over the world. For instance, the Belle collaboration (http://belle.kek.jp/), with 54 research groups (institutional members), is one such international undertaking. The University of Melbourne's Experimental Particle Physics group (http://www.ph.unimelb.edu.au/epp/) is a member of this collaboration. An EHEP collaboration is established to find answers for a narrow range of questions. For instance, the Belle collaboration is set up to study the nature of Charge-Parity symmetry violation, which the physicists believe may explain the dominance of matter over anti-matter in the universe. There are other contemporary collaborations, such as the CLEO collaboration (http://www.lns.cornell.edu/public/CLEO/) with 25 research groups and the BaBar collaboration with 77 research groups (http://www.slac.stanford.edu/BFROOT/), are involved in a similar study [1].

The research groups within a collaboration analyse the huge sets of data produced in an experiment, using various analysis techniques. The analyses attempt to discern the hidden pattern inside the data sets. Typically, the results of the analyses are communicated to fellow researchers in the form of pre-prints and research notes and is stored in the collaboration's publication archive, like http://belle.kek.jp/belle/publications/. This kind of information sharing leads to scientific productivity and trust. A research group is free to verify and extend the analysis work of another.

While the information about the experiment and its data is evident to all members of the collaboration, the research group that performed the analyses holds the complete knowledge about the analyses. Most EHEP publications do not provide detail description about the analyses performed. 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 publications mainly highlight the results of the analyses that account for the observed phenomena.

The Experimental analyses described in this fashion, with publication bogged down with tacit knowledge are prone to be misunderstood, particularly by new researchers or 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.

3. EXPLICATING THE EXPERIMENTAL ANALYSES

It is not difficult to see that the problem in the scenario described above could be traced to lack of structure and semantics in the published analysis description documents. Debugging an experimental analysis described by authors who profess somewhat different ontological commitment about the domain is indeed a daunting task. We believe this misinterpretation problem 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. A systematic elaboration of the analyses would allow for a clear and detailed description of the content. Publishing the analyses with annotations that further enrich its description can ensure optimal exchange of information between researchers within a collaboration.

Moreover, these machine-readable ontologies can also be utilised to describe analysis jobs. The formal specification of description can be interpreted runtime by *analyser* agents to perform the required data analyses.

The EHEP ontologies can also 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.

This opportunity to embark upon an innovative way of handling scientific information generated within an EHEP collaboration is illustrated in **Fig. 1**. It affirms the belief that the next generation web can indeed change the way scientific knowledge is produced and shared, as envisaged by Berners-Lee and Hendler [5].

4. CREATING THE EHEP ONTOLOGIES

The EHEP ontologies will be developed to be reused across different applications as depicted in section 3. The ontologies emphasise the formal semantics and capture the intrinsic structure of the domain embodied as concepts, relations and axioms. The creation of the EHEP ontologies is carried out in stages. First of all, there is a need for the ontologists to attain sufficient level of literacy in the EHEP domain to facilitate the impending knowledge acquisition task. Initial discussion with the EHEP physicists and related literature review enabled us to identify the main domain concepts in a typical EHEP experimental analysis. These concepts will become the ‘hooks’ in the skeletal EHEP knowledge model.

Next, each of these ‘hook’ concepts is expanded systematically, as sub-models of the EHEP domain. These models are in essence, taxonomies of defined concepts with their roles (properties) restricted. The knowledge models are 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.

We are developing these models using a Frame-based tool, called Protégé-2000 [2]. Frames provide an object view of the world and an intuitive modelling style. In spite of some modelling limitations, Protégé-2000 still is a useful interaction tool for eliciting knowledge from the EHEP physicists.

A parallel activity undertaken during this time is the formulation of a set of competency questions that outline the competence of the EHEP ontologies. The regularly updated competency questions effectually guide the acquisition of the correct domain knowledge for the models.

In short, the development of the knowledge models follows an evolutionary development cycle, which also encompasses the model validation, verification and refinement. This is part of our ongoing work.

Finally, the completed models will be formalised as EHEP ontologies. We intend to implement the ontologies in DAML+OIL [4], a Frame and Descriptive Logic integrated ontology language, which is set to be the standard semantic markup language for web resources.

5. OUR MAIN RESEARCH ISSUES

The EHEP ontologies will provide the framework for communication, integration and sharing of resources among the distributed research groups. It is the foundation for the web services that will be enacted for the EHEP community. In the process, this research project is set to investigate the two key issues:

- How well can we express the domain knowledge pertaining to the EHEP experimental analyses in a natural way (mirroring the real world semantics)?
- There is a concomitant need to mark-up data and information regarding experimental context in the scientific documents, before it can be used as knowledge. How can we facilitate the annotation of the EHEP collaboration documents?

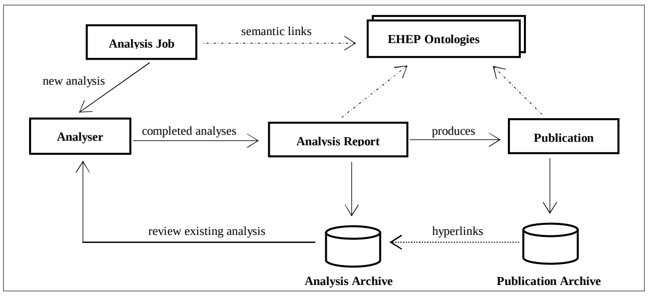

Fig. 1. Handling EHEP collaboration documents. Researchers prepare and deliver the semantically marked up analysis reports and publications, which can be archived and referred during subsequent experimental analysis. The content of the archives can also be searched more productively using precisely defined queries. Jobs described using the ontological terms can be processed directly by the agent analyser

6. REFERENCES

[1] Board on Physics and Astronomy, National Research Council, USA: Elementary-Particle Physics - Revealing the Secrets of Energy and Matter, National Academy Press, Washington D.C., 1998.

[2] Grosso, W.E., Eriksson, H. *et. el.*: Knowledge Modelling at the Millennium (The Design and Evolution of Protégé-2000), In the Proceedings of KAW99, 1999.

[3] Gruber, T: A Translation Approach to Portable Ontologies, Knowledge Acquisition, Vol. 5, No. 2 (1993), 199-220.

[4] Hendler, J. & McGuiness D: The DARPA Agent Markup Language, IEEE Intelligent Systems, Vol. 15, No. 6 (2000), 72-73.

[5] Lee, T.B & Hendler J: Scientific Publishing on the Semantic Web, Nature, April 2001.