On Extending the Knowledge Sharing Practices in Scientific Collaboration to the Semantic Web

Muthukkaruppan Annamalai[a], Leon Sterling[b]

[a]*Faculty of Information Technology and Quantitative Sciences*
MARA University of Technology, 40450 Shah Alam, Selangor, Malaysia
Tel : +60 3 5521 1161, Fax : +60 3 5543 5501
E-mail : mk@tmsk.uitm.edu.my

[b]*Department of Computer Science and Software Engineering*
The University of Melbourne, Victoria 3010, Australia
Tel : +61 3 8344 1404, Fax : +61 3 9348 1184
E-mail : leon@cs.mu.oz.au

ABSTRACT

The internet has greatly facilitated scientific collaborations to efficaciously share their analytic knowledge. The impending semantic web promises to dispense with some of the human effort by making the shared knowledge accessible to machines on a semantic basis; giving rise to the connotation of webifying artificial intelligence (AI) for doing reasoning of science, if unchecked, could turn into another pipe-dream. In practise, as exemplified in this paper, it requires considerable effort to put the shared knowledge in common. Therefore, we stress the need to adopt a realistic attitude when extending the knowledge sharing practices in scientific collaboration to the semantic web.

Drawing ideas from previous studies in academic communities, supported by our analysis about knowledge sharing in the case study domain of Experimental High-Energy Physics (EHEP), and acknowledging the limitation of AI, we advocate pursuing a pragmatic approach that emphasises the sharing of the analytic knowledge of a domain which is employed in computation, not cognition. Further, we highlight that it is suffice to focus on the directively and paradigmatically shared knowledge in a scientific collaboration.

Keywords
Knowledge sharing, Knowledge modelling, Scientific collaboration, Semantic web

1.0 INTRODUCTION

A scientific collaboration is established to investigate complex phenomena that are beyond the capabilities of an individual scientist or group. Confederations of scientific groups with common research goals pool their expertise and resources together to undertake collaborative scientific effort. Today, large scale scientific collaborations encompass widely distributed scientific groups involving hundreds of institutions across the globe in research areas such as Human Genome, Oceanography, Astronomy and High-Energy Physics.

In general, an experiment of a scientific collaboration is set up centrally and is staggeringly data intensive, resulting in large data collections. The data is methodically captured and stored according to predetermined experimental and observation methods. After that, each scientific group is free to plan and analyse the shared data in fitting manner using various analysis methods and techniques known to them.

An important component of collaborative scientific research is the sharing analytic knowledge and their related findings. The internet has become the major vehicle for distributed groups in scientific collaborations to share knowledge related to their research quickly (Atkins et. el., 2003). Yet, there is still much human mediation to correctly interpret and reuse the shared knowledge.

The next generation web, called the semantic web (Berners-Lee, 1998), the eScience research initiatives (Hey & Trefethen, 2002) and the emerging grid (Foster, 2002) promise to dispense with some of the human effort by making the shared knowledge accessible to machines on a semantic basis to solve scientific problems operating on a global scale. Since it is difficult to bring the machine to make sense about the physical world, the objective of the semantic web is to represent the human knowledge formally with the aid of ontologies or knowledge models (Jacob, 2003), enabling machines to interpret and use the knowledge.

Firstly, the sharing of the analytic knowledge in scientific collaborations on the semantic web has yet to receive proper attention. Secondly, since the semantic web concepts as logic, knowledge modelling and representation overlap with that of AI, exemplary semantic web initiatives highlight the attempt to webify AI for doing reasoning of science as envisaged in the motivating articles on science and semantic web as

Berners-Lee & Hendler (2001), Hendler (2003), Shadbolt et. al. (2006), and also attempts to heed such call as in King et. al. (2004). The paper addresses this concern and stresses the need to adopt a more realistic attitude when extending the existing knowledge sharing practices in scientific collaborations to the semantic web; so as not to turn the endeavour into hype. In reality, formalising the knowledge for scientific reasoning by machine requires modelling of "deep knowledge"; which is a daunting task. Besides, the expressivity and scalability afforded by contemporary semantic web languages is much restricted (Goble, 2005). Consequently, a question that arises is: "*What aspects of the analytic knowledge shared in a scientific collaboration do we need to model* and *why?*"

Our analysis examines issues related to knowledge sharing in the case study domain of EHEP, where there are many large scientific collaborations (Board of Physics and Astronomy, 1998). Consequently, the examples in this paper will be derived from cases in EHEP. We are convinced that our findings are relevant to other scientific collaborations that are about to embark on the semantic web.

2.0 THE EHEP SCIENTIFIC COLLABORATION: A CASE STUDY

EHEP is dominated by large scientific collaborations, with membership from all over the world. In this field of our case study, contemporary scientific collaborations such as the Belle collaboration with fifty-four institutional members (http://belle.kek.jp/belle/) and BaBar collaboration with seventy-seven members (http://www.slac.stanford.edu/BFROOT/) are actively involved in the study of the nature of Charge-Parity symmetry violation, which the physicists believe may hold vital clues to explain the dominance of matter over anti-matter in the universe. The symmetry violation is evident in certain rarely occurring B meson particle decay events.

Whilst a B event of interest occurs scarcely, the noise, i.e., the background events that accompany the signal, is enormous. A particular data analysis in EHEP attempts to recognise interesting decay patterns existing in the huge experimental data sets by systematically removing superfluous noise. Events for acceptance or rejection are characterised by applying quantitative restrictions on various distributions involving event selection variables, which are typically referred to as cuts.

The recording and sharing of the analytic knowledge is an integral part of collaborative scientific research because it generates the essential shared understanding to enable researchers to collaborate. For example, in the Belle collaboration, these publications are often referred to as Belle Notes and are maintained in the collaboration's secured online repository (http://belle.kek.jp/secured/belle_note/). The sharing of knowledge helps to validate and verify previous findings and as well as to advance their research, leading to scientific productivity and trust.

3.0 KNOWLEDGE SHARING IN ACA-DEMIC COMMUNITIES

Studies on the scholarly network in sciences show that tightly bound scientific researchers exhibit a high level of mutual dependency upon analytic knowledge produced in works of peer researchers and ensuing analyses are often built on earlier results to pursue their common research goals (Fuchs, 1992; Whitley, 2000). The existence of functional dependencies among peer researchers is demonstrated by their adherence to common work practice and competence standards. Social cognitive theory also contends that people actively organise and regulate their actions largely on the basis of gained experience (Bandura, 1986). As a result, the cognitive behaviour of the researchers is to a large extent founded on socially and technically shaped analytic experiences in scientific collaboration.

Sharing of analytic knowledge supports learning and constant experimentation in scientific collaboration. The changing understanding contributes to the incremental knowledge and development of innovative analytical techniques and methods. Not surprisingly, much importance is placed upon the prompt dissemination of findings to guide future analytical and decision processes, as apparent in the establishment of in-house communication channels, such as online repositories and pre-print systems in scientific collaborations. Such mode of dissemination of analytic knowledge underscores the *directive*, *paradigmatic* and *strategic* knowledge sharing practices on which the learning process in scientific collaborations is based. This characterisation of the types of knowledge sharing practices in scientific collaborations roughly follows the example of Talja (2002).

Talja introduced the concepts of *directive*, *paradigmatic*, *strategic* and *social* information sharing to describe the different levels of information sharing in the context of information seeking and use in an academic organisation. According to Talja, *directive* sharing is recognised in the information sharing between a teacher and a student; *paradigmatic* sharing is perceived as a course to establish a novel research approach in a research; *strategic* sharing is a deliberate strategy of maximising efficiency in a project team; and *social* sharing as a means of establishing relationship among academician with diverse research interest.

Adopting and adapting Talja's conception to sharing of analytic knowledge in scientific collaboration, we rephrase **directive knowledge sharing** as apparent in the sharing of centrally controlled information (recorded knowledge) about the experiment (e.g. supporting scientific theories) and empirical knowledge about

material being examined (e.g. properties of sub-atomic particles) underwritten by high fidelity and often maintained by bodies of authority in the subject domain. The *directive* sharing coincides with the practice of dissemination of ratified knowledge in a domain. The **paradigmatic knowledge sharing** involving distinguishable analytical techniques and methods helps to establish a shared understanding that can serve as a prototypical analysis approach within a scientific collaboration. We view the **strategic knowledge sharing** as evident in the sharing of exclusive analytic resources aimed at maximising the efficiency of its consumption to make rapid progress in a closely-knitted institutional research group, i.e., in the same vein as Talja's conception. **Social knowledge sharing** as part of community building effort is *not* obvious in scientific collaborations that are comprised of researchers from a single discipline and pursuing common research goals.

The *strategic*, *paradigmatic* and *directive* knowledge sharing modes reflects the knowledge flow within an expanding community of interactions along the ontological dimension in the Nonaka's (1994) Spiral model.

In what follows, we shed some light on the different kinds of analytic knowledge resources shared in our case study domain of EHEP, organised according to their level of sharing in the scientific community. We broadly consider three categories of knowledge sharing, namely *directive* sharing across the domain, *paradigmatic* sharing within a particular scientific collaboration, and *strategic* sharing restricted to a research group.

3.1 Directive Sharing of Factual and Empirical Knowledge in the EHEP Domain

We have identified two types of shared knowledge resources in this category. The dissemination of this ratified knowledge is considered to be widely applicable technical reference. Its organisation is set apart in determinate ways from context of use.

I. Theoretical account supporting EHEP research findings are published in books, scientific literature and journals of interest to particle physics.

II. Certified empirical knowledge regarding subatomic particles, the characterisation of particle dynamics and intrinsic properties of these particles. This kind of systematised knowledge is often shared in the form of technical references such as the Review of Particle Physics handbook (2004). Other organised knowledge resources of value to the HEP communities include information about fundamental physical constants (http://phy-sics.nist.gov/cuu/Constants/), a standard list of HEP Monte Carlo particle numbers and numbering schema (http://www-pat.fnal.gov/ stdhep.html) to facilitate interfacing with tools used in particle physics, and natural units of measurement.

3.2 Paradigmatic Sharing of Analytic Knowledge in a Scientific Collaboration (e.g. Belle Collaboration)

We have identified three types of shared knowledge resources in this category. This form of knowledge sharing takes place to support researchers of a scientific collaboration with common research concerns. This translates to sharing of pertinent knowledge, either in packaged or recorded form, involved in the mechanism of their common and purposeful task activities.

I. The experimental data analyses undertaken by EHEP researchers are directed to precisely measure the crucial behavioural attributes associated with subatomic particles. The analysis performed on the data and the empirical findings are readily shared among peer researchers in the form of preprints and research notes.

II. The sharing of packaged knowledge in the form of statistical programs and modules supported by statistical analysis tools such as PAW (http://paw.web.cern.ch/paw/) and particle path simulation tools such as GEANT (http://www-asd.web.cern.ch/wwwasd/geant/) is common in EHEP scientific collaboration.

III. Knowledge about the organisation of the real data captured by detectors is comprised of the data model and supporting provenance information. The sharing of this knowledge is important for proper access and handling of the related experiment data, often stored in distributed data repositories.

3.3 Strategic Sharing of Analytic Knowledge in a Research Group (e.g. The University's Experimental Particle Physics Group)

We have also identified three types of shared knowledge resources in this category. The *strategic* sharing practice thrives in local research group largely because of the strong social interaction and trust that exist within the group.

I. Recorded knowledge, which includes analysis documentation, reference notes and work files (e.g. event profiles and statistical summaries) generated in the course of data analysis.

II. Packaged knowledge such as source and object software modules (e.g. event analyser, simulation data generation program) for data analysis and simulation.

III. Knowledge about the organisation of locally maintained simulation and skimmed event data files.

Although we have highlighted a variety of knowledge resources shared within a scientific collaboration, we advocate that we initially focus on extending the existing publishing model perspective to share the recorded analytic knowledge on the semantic web. The sharing of packaged knowledge can come later.

4.0 ANALYTIC AND BACKGROUND KNOWLEDGE

Collaborative scientific research is task-oriented. The task undertaken by data intensive scientific collaboration is the systematic analysis of huge data collection under their purview. The analytic knowledge that emanates from the data analysis task is shared in the scientific collaboration. In order to characterise the structure of the analytic knowledge involved in carrying out a task, we asked the following questions: *how* the task is performed and *what* knowledge is utilised (or generated), *when* the knowledge is needed and *why* the knowledge is admissible. The response to the *what* and *why* questions expose the underlying domain knowledge structure, whereas the *how* and *when* questions appeal to the task knowledge structure. Consequently, we recognised the corresponding types of background knowledge utilised in data analysis were identified, which includes, but is not necessarily limited to *what*, *why*, *how* and *when* knowledge.

We provide below an extended description about the four types of background knowledge in our case study. Whether these apparent differences correctly categorise the background knowledge is not our immediate concern. The terminology is only suggestive, however will prove useful in the ensuing discussion.

4.1 *How* knowledge

This describes a task activity, i.e., a structured description of a set of steps to accomplish a specific task, data analysis in our case. This task knowledge is derived based on community of practice in solving similar problems and understanding the limitations of specific solution. For example, the *event analysis* task is a common task undertaken by the EHEP community, whose subtasks include: *Determination of signal and background events, Determination of event selection criteria* and *Evaluation of signal efficiency*.

4.2 *What* knowledge

This provides a description of terminologies practicable in the scientific domain, in terms of which abstract entities are expressed with precision and brevity. For example, the B^0 is defined as a neutral meson particle comprised of a pair of d and d^- quarks, with zero electric charge, has a nominal mass of 5279 MeV/c^2, and so on. This knowledge about the subatomic particle B^0 helps to distinguish it from another particle.

The *what* knowledge also clarifies the analytic knowledge that serves common needs for event analysis in the EHEP domain such as feasible B^0 decay channel like B^0 ? ?$^0p^0$; ?^{0} ? p^+p^-; p^0 ? ? ?, and that B^0 decays in this mode with branching fraction upper limit of 5.3 x 10^{-6}, the description of cuts utilised to reconstruct the decay particles, and so on.

4.3 *When* knowledge

This makes reference to a particular functional subtask in a task activity. It identifies the material context for the analytic knowledge, i.e., when the pieces of analytic knowledge is produced or applied; in a way indicating the intention for its use. For example, in a EHEP data analysis, the *when* contextualises a set of cuts on track attributes that reconstructs the B^0 candidate particle to the *signal event selection* task.

4.4 *Why* knowledge

This supplies the justification or causal reasons to support certain claims about a particular subject matter. In reality, the justification is linked to a body of cognition acquired by combining and extending knowledge gained from scientific sources (principles and theories), experience, reasoning and insights, which are embodied in community of practice, and sometimes are subjective and situation-dependent.

For example, a *cut* on the *Likelihood of pion over kaon*: *L{p/?} > 0.6* is applied to identify *pion*, i.e., $p^+$and p^- particles in the B^0 ? ?$^0p^0$ decay channel. The reason *why* this cut is necessary is because of the need to ensure that no *pion* is misidentified for *kaon* in the generic background decays such as *B ? Dp* types, where *D ? ?* p^+p^-. The question as to *why-not* a tighter cut is employed with a value higher than *0.6* is because it has been found to be statistically most effective when tested on appropriate Monte Carlo simulation data; a tighter cut can possibly lead to the signal *pion* be rejected. The reason *why* the above cut on the distribution of *L{p/?}* works can be traced to previous findings of the scientific community that show that this cut is able to identify *pion* particles with high efficiency in the momentum region of *pion* between 500 MeV/c and 2000 MeV/c in this kind of analysis. The question as to *what-if* we also need to identify *pion* particles with momentum outside this region may have to be dealt in the computation of error analysis.

5.0 DISCUSSION AND CONCLUSION

The previous section highlighted some different types of background knowledge that helps to clarify the shared analytic knowledge in scientific collaboration. The *what* and *why* knowledge are imperative when considering the modelling of the domain, while the *how* and *when* knowledge distinctively characterise the task. Relevant knowledge models must be developed to make this background knowledge explicit.

In practise, the knowledge modelling effort is demanding and time consuming. Therefore, the incentive must be commensurate with the level of effort that must be

expended for the scientific collaboration to view it as worthwhile. For example, while it may be encouraging to focus on the modelling of analytic knowledge that researchers need for cognition, which may be of significant use to some researchers, but we must weigh the effort it takes to put such analytic knowledge in common. Hence, it is necessary to have a clear understanding of the purpose for developing a certain kind of knowledge resource collection for a scientific collaboration. We underscore the importance of adopting a realistic attitude when extending the knowledge sharing practices in scientific collaboration to the semantic web. We propose two dimensions for consideration: the scope of knowledge sharing and the types of knowledge to model.

5.1 Knowledge Sharing Scope

On the first dimension, we want to emphasise that a critical mass of users are required to justify the modelling effort. There is little incentive to semantically-enrich the shared knowledge resources utilised only by a small group of researchers, as the potential of use and reuse is limited. Therefore, we advocate focussing on domain-wide *directively* shared knowledge, and collaboration-wide *paradigmatically* shared knowledge.

The *directively* shared analytic knowledge provide a source of ratified reference knowledge of high integrity that are incorporated into an analysis, while, the *paradigmatically* shared analytic knowledge can serve as intellectual resources for emulating and extending previous analyses.

5.2 Types of Knowledge

On the second dimension, we recognise the modelling of useful knowledge to distinctively characterise the task and the domain.

In this regards, a task (*how*) knowledge structure serve as the basis for streamline communication between researchers and effective connection between researchers and their computational tools on the semantic web. The task structure also helps to indicate the context (*when*) of production or use of the analytic knowledge to support a task activity.

As for domain knowledge modelling, we need to represent both *general* and *purposive* knowledge. While it might be possible to rely on the broadly disseminated knowledge about the domain to model the *general* domain knowledge, the modelling of the *purposive* domain knowledge focuses on the integral part of resources related to certain analytic tasks that are either computative or cognitive in nature. The description of *purposive* knowledge applied in a task is sometimes referred to as "support knowledge" (Bylander & Chandrasekaran, 1988; Clancey, 1987), which represents the understanding of domain when performing that task.

For a computative task, the *purposive* knowledge merely relies upon the semantic content as *what* knowledge (c.f. glossary of technical terms) to correctly interpret the `instructive' input and carry out the operation effectively. On the other hand, a cognitive task additionally requires *purposive* knowledge about causal process with intentional characteristics, i.e., *why* knowledge in the form of abstraction to support reasoning and decision making.

The modelling of *why* knowledge for cognitive utilisation is the principle of problem solving in AI. The knowledge engineering approaches such as CommonKADS (Schreiber et. al., 2000) recommends acquisition of this knowledge based upon a predetermined inferencing strategy for specific cognitive tasks, but is still dependent upon the availability of potent inferential knowledge to instantiate it. The issue is in the development of efficacious *why* knowledge exhibiting high-level intent of cognitive significance to either influence or determine the cognitive behaviour. Developing the *why* knowledge model takes for granted that we could provide the knowledge that can be utilised to support cognition. In reality, fulfilling this knowledge need is not easy for the two reasons mentioned below.

Firstly, in regards to a machine's inability to learn as effectively as humans do, the formidable task of acquiring and modelling of the *why* knowledge and keeping the knowledge resources up to date becomes the sole responsibility of the modeller.

Secondly, the resources to develop the *why* knowledge model is unfortunately limited by quality and accessibility, particularly when the knowledge to be represented is complex and dynamic such as is the case in scientific domains, which could leave us quite stuck. We not only require knowledge about the expertise involved in a real situation, but also the extent and origin of the many different information pertaining to it, as well as the reliability of that expert knowledge in the context of use. This means we also have to deal with much of the reasoning uncertainties in the form of knowledge of purpose, practice and performance in situation (Lakoff & Johnson, 1999). For example, it became apparent from the discussion in the previous section that providing *why* knowledge demands consideration not just limiting to straightforward causal association in response to *why* questions, but also to factor in the knowledge that can respond to *what-if* and *why-not* questions that arise in the causal understanding.

Because of the bottleneck in the acquisition and building up of the *why* knowledge content, we believe this part of AI subset that aims to capture `intelligence' and equipping machines for doing reasoning of science is not

immediately realisable. So, we argue that the development of *why* knowledge to support cognition is an unpromising path to go for now.

A potentially fruitful path to pursue is to better facilitate human understanding, as well as to expedite tasks as data analysis by making it easier for formal tools and intermediaries to interpret the specifications of analysis, clarified using precise and appropriate metadata (*what* knowledge) to support knowledge lean tasks as retrieval, integration and computation in order to compare the experimental findings and validate the results. In comparison to *why* knowledge, the *what* know-ledge is amenable to development due to their abiding nature in a scientific domain and ease of their accessibility and maintainability.

In summing, we perceive the *general* and *purposive* domain knowledge models as the determinant of the operationality of tasks commonly engaged by researchers in scientific collaboration. In the context of knowledge sharing in a scientific collaboration, the *general* and *purposive* domain knowledge correspond to the *directively* and *paradigmatically* shared knowledge in a scientific collaboration, respectively.

REFERENCES

Atkins, D., Droegemeier, K., Feldman, S., Garcia-Molina, H., Klein, M. & Messina, P. (2003). Revolutionizing Science and Engineering through Cyberinfrastructure. Technical Report. NSF Blue-Ribbon Advisory Panel.

Bandura, A. (1986). Social Foundations of Thought and Action: A Social Cognitive Theory. Prentice Hall, New York, 324–.

Berners-Lee, T. (1998). Semantic Web Road Map. http://www.w3.org/Design-Issues/Semantic.html/

Berners-Lee, T. & Hendler, J. (2001). Scientific Publishing on the Semantic Web. Nature.

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

Bylander, T. & Chandrasekaran, B. (1988). Generic Tasks in Knowledge-based Reasoning: The Right Level of Abstraction. Knowledge Acquisition for Knowledge-based Systems, 1, 65–77.

Clancey, W (1987). Knowledge-based Tutoring: The GUIDON Program. MIT Press, Reading, MA.

Foster, I. (2002). The Grid: A New Infrastructure for 21st. Century Science. Physics Today.

Fuchs, S. (1992). The Professional Quest for Truth: A Social Theory of Science and Knowledge. SUNY Press, New York.

Goble, C. (2005). Using the Semantic Web for e-Science: Inspiration, Incubation, Irritation. LNCS 3729, 1–3.

Hendler, J. (2003). Science and the Semantic Web. Science, 299.

Hey, T. & Trefethen, A. E. (2002). The UK e-Science Core Programme and the Grid. Future Generation Computer Systems, 18, 8, 1017–1031.

Jacob, E. K. (2003). Ontologies and the Semantic Web. Bulletin of the American Society for Information Science and Technology, 29, 4, 19–22.

King R.D, Whelan K.E, Jones M.F, Reiser P.G.K & Bryant C.H. (2004). Functional Genomics Hypothesis Generation by a Robot Scientist. Nature. 427, 247–252.

Lakoff, G. & Johnson, M. (1999): Philosophy in the Flesh: The Embodied Mind and its Challenge to Western Thought. Basic Books, New York.

Nonaka, I. (1994). A Dynamic Theory of Organisational Knowledge Creation. Organisational Science, 5, 1, 14–37.

Review of Particle Physics (2004). Physics Letters B, 592.

Schreiber, G., Akkermans, H., Anjewierden, A., de Hoog, R., Shadbolt, N., van de Velde, W. & Wielinga, B. (2000). Knowledge Engineering and Management – The CommonKADS Methodology. MIT Press, Cambridge, MA.

Shadbolt, N., Berners-Lee, T. & Hall, W. (2006). The Semantic Web Revisited. IEEE Intelligent Systems 21, 3.

Talja, S. (2002). Information Sharing in Academic Communities: Types and Levels of Collaboration in Information Seeking and Use. New Review of Information Behaviour Research, 3, 143–159.

Whitley, R. (2000). The Intellectual and Social Organisation of the Sciences. Second edition, Clarendon Press, Oxford.