Between too informal and too formal

Dr Philippe Martin, **Dr Mike Eboueya**, **Dr Jun Jo** **and** **Dr Lorna Uden**

Keywords: domain modelling, ontology-building, semantic modelling/webs,
knowledge sharing/representation/retrieval/server

Extended Abstract

Very few information repositories, especially corporate memories, are formal knowledge bases (KBs) since, despite its benefits, formal and scalable knowledge modelling is inherently a difficult and time-consuming exercise that current generic KB systems (KBSs) still do not guide well. Instead, repositories are most often composed of informal documents that are independently created by people (typically by publishing a Web document or sending an email to a mailing list). This approach is simple but the well-known drawback is that it is then often difficult to retrieve or compare information because (i) the various needed pieces of information are scattered in many documents and expressed in different ways and at often inadequate levels of details, (ii) these pieces of information cannot be automatically organised into a semantic network by any current (or even currently foreseeable) natural language understanding (NLU) technique. The use of cooperatively edited informal documents (as in wikis) helps to reduce the scattering of information but introduces new problems and does not in itself lead to better or sufficient structuring of the information. Structured documents (e.g., documents following an XML schema), databases and application-oriented interfaces enforce some structure but (i) they also often restrict what can be entered even when they provide "free text" entry fields, and (ii) the semantic of the prescribed structure is most often left implicit and is insufficient to be used by KBSs. The occasional use of semantic relations (as in semantic wikis) or of metadata (typically, RDF metadata; especially those automatically generated by tools during document edition), are also insufficient for automated reasoning purposes and hence for "knowledge retrieval", i.e., precision-oriented information retrieval. The use of poorly structured, graphical and overly permissive semi-formal notations such as those used for Topic Maps often lead to information that are more difficult to understand, retrieve and exploit than when regular informal sentences are used. Controlled languages, i.e., semi-formal languages that look like natural languages but have a restricted syntax or a restricted vocabulary, are often seen as good compromises between formal and informal languages but (i) they do not scale (i.e., in the general case, they are not expressive and formal enough or become too complex to use), (ii) they are often more structured, precise, normalised and readable ways to express knowledge (for example, partOf or generalization hierarchies).

Research in knowledge representation has focused on reasoning and therefore not on creating *notations that are both very readable and expressive*, although John Sowa had this in mind when creating the Conceptual Graph linear and graphical notations (Sowa, 1984) and his relative success is what still makes much of appeal of the CG formalism and approach (a related reason is that, being higher level than other notations, the CG notations lead to more "normalised" knowledge representations which ease knowledge sharing and retrieval). It is however possible to go further in terms of "high-level-ness", i.e., readability (concision and/or intuitiveness) and normalising effect. (Martin, 2002) did so by creating the Frame-CG (FCG), Formalized-English (FE) and For-Taxonomy (FT) notations. FT is for representing relations between non-quantified types or individuals (e.g., statements). FE looks like some pidgin English but is structurally equivalent to FCG which is an extremely concise notation that includes constructs for extended quantifiers, contexts, functions and various interpretations for sets (hence, it is semantically equivalent to KIF). FE is quite verbose and hence is not adequate for really building or browsing a reasonably complex KB but it can be shown to anyone. Hence, it can for example be used for showing the various interpretations that a NLU parser makes of a sentence expressed in a natural or controlled language and then let the user select the right interpretation or precise its sentence. These notations are regularly extended, e.g., FT has just been extended for easing the (re)presentation of structured discussions, which makes it a good (and often more expressive) alternative to the notations used in argumentation systems (e.g., IBIS). (The article will provide examples). The authors of typed hypertext systems and argumentation systems have often argued that adopting a knowledge-based approach (e.g., allowing the users to update an ontology instead of restricting them to use a few predefined relation types and concept types) would scare many of the users and that this would prevent a wide adoption of their tools. However, none of these systems achieved wide adoption. This may be attributed to the fact that the restrictions deeply limited what could be done with their tools, and hence their interest and applicability. Furthermore, the restrictions also led to biases representations and complex turnarounds. Similarly, it is a mistake to restrict the expressivity of a *general* knowledge representation language since choices about how to handle the completeness, decidability and efficiency issues, or how to handle elements such as sets and modalities, are application-dependant (e.g., for some knowledge retrieval or filtering purposes, efficient graph-matching procedures that ignore the detailed semantics of certain elements can be used, while for other purposes exploiting all the details is essential). To conclude, for readability reasons and to support various kinds of knowledge entering or views on the knowledge, various notations should be supported but precision and normalisation should always be encouraged. The article will also discuss the need for *various ways of querying and comparing knowledge* and will illustrate the proposed new approaches.

A minimalist *knowledge sharing strategy* is the one envisaged by the W3C for the "Semantic Web": many small documents (containing ontologies or knowledge representations) more or less independently developed and thus partially redundant, competing and very loosely interconnected. This approach is inadequate for a knowledge repository. Indeed, (i) finding the relevant ontologies, choosing between them and combining them requires commonsense and a lot of background knowledge (and hence is difficult and sub-optimal even for a knowledge engineer), (ii) a user cannot simply add one concept or statement "at the right place" (she has to create an ontology and make connections to concepts or statements of many other ontologies), and she is not guided by a large ontology (with a system exploiting it) into providing precise concepts and statements that complement existing ones and are more easily found and re-used, and (iii) the result is often more or less lost to others and increases the amount of information to search. There now are many tools to extract knowledge from texts/databases or align concepts of different ontologies; they are very imperfect although they can be sufficient for certain applications. Hence, a knowledge repository, whether it is implemented as a peer-to-peer network or as a knowledge server that let its users edit a shared KB, requires *protocols in order to maximise conceptual relations* between the objects (categories and statements) created by the various users, and hence permit the comparison of these objects and remove redundancies. The creator of each object must also be represented in order to permit filtering on knowledge sources and to quantify the *popularity and originality of each contribution and contributor*. This article will summarize our approach (Martin, 2005), extend it slightly, and will compare it with other approaches.

To permit knowledge comparison and retrieval, ease and guide knowledge entering, and support automatic knowledge extraction, a knowledge repository should be built upon a large initial KB. This article will summarize various ontology integration and experiments (including very recent ones) that we have done to build backbones of knowledge repositories.

To conclude, this article will synthesize and refined the results of various works that we have done to permit information repositories to be knowledge repositories.

References

Martin Ph. (2002). *Knowledge representation in CGLF, CGIF, KIF, Frame-CG and Formalized-English.* Proceedings of ICCS 2002, 10th International Conference on Conceptual Structures (Springer Verlag, LNAI 2393, pp. 77-91), Borovets, Bulgaria, July 15-19, 2002.

Martin Ph., Blumenstein M. & Deer P. (2005). *Toward cooperatively-built knowledge repositories*. Proceedings of ICCS 2005, 13th International Conference on Conceptual Structures, (Springer Verlag, LNAI 3596, pp. 411-424), Kassel, Germany, July 18-22, 2005.

Sowa J.F. (1984). Conceptual Structures: Information Processing in Mind and Machine. Addison-Wesley, Reading, MA, 1984.