Allen, R.B., Rich Semantics and Direct Representation for Digital Collections. ACM/IEEE JCDL, 2017, 348-349.

Rich Semantics and Direct Representation for Digital Collections

Robert B. Allen
Yonsei University
Seoul Korea

rballen@yonsei.ac.kr

ABSTRACT

It is now possible to envision the close integration of rich knowledge structures and knowledgebases with digital libraries. Yet, there are many challenges to the implementation of this vision. Chief among these is finding comprehensive and rigorous, but also flexible, representations. Such representations need to go beyond semantics strictly construed to include discourse, the evolution of knowledge, and support for alternate explanations. In this endeavor, there are many traditions to draw from such as LIS, linguistics, programming languages, philosophy, jurisprudence, sociology, and systems analysis. While the most obvious application is to develop highly-structured scientific research reports, rich semantic information organization could be applied to areas including law, history, and biography. We propose a community-wide exploration of these issues and the development of a new generation of digital libraries.

Keywords

Causation, Discourse, Events, Frames, Highly-Structured Repositories, History, Models, Ontologies, Programming Languages, Scholarly Resources, Science, Systems

1. INTRODUCTION

Rich semantics can support detailed information organization for the contents of documents, across documents, and even across resources in different modalities. In its strongest form, direct representations consist of complex material composed entirely of knowledge structures. The advantages of large collections with rich semantics and direct representation includes access and integration of knowledge resources. Highly-structured representations could eliminate the ambiguity of text and the need to implement complex text processing. Eventually, we hope to represent and coordinate large collections across different disciplines in which knowledge is continually evolving and we hope to support a broad range of services such as providing alternative explanations and interpretations. However, there are a great many challenges to this effort.

2. RICH SEMANTIC REPRESENTATIONS

While there is now considerable awareness of the importance of semantics, work on semantics is still fragmented. Rich semantics goes beyond simple models for linked data such as those using RDF-based triples and beyond ad hoc ontologies. Potentially, rich semantic frameworks would include complex entities, dynamic models, schemas, systems, and descriptive programs.

Some approaches to semantics are more relevant than others to information organization. For instance, while automated inference can be useful, the adequacy of semantic systems for description seems more important than inference. Similarly, in this context, we are not primarily focused on the semantics of human cognition and natural language.

Ontologies are an obvious foundation for our effort but there is a wide variety in the rigor and scope of ontologies. Systems of reference ontologies based on rigorous upper ontology seem most promising. In addition to the structure provided by the upper ontology, we also need rich knowledgebases of parts, qualities, and mechanisms associated with the objects which are defined by the *is_a* hierarchies.

Moreover, the current generation of ontologies is primarily based on triples. For the representation of digital collections, we believe that rich semantic structures such as systems, models, simulations, events, and frames are needed. It is surprising to us that there has been relatively attention paid to the coordination of ontologies with programming language semantics. Indeed, many ontologies seem stronger in the representation of static entities (e.g., continuants, endurants) than dynamic entities (e.g., occurrents, perdurants) while the opposite is true of programming languages. Beyond the need for coordination with programming-language structures, it also seems useful to develop connections between rich semantics and methodologies such as systems analysis, object-oriented design, conceptual modeling, composable simulations, and business process engineering. For instance, systems analysis might be applied to describing the activities in a town from a historical newspaper.

3. DISCOURSE AND EPISTEMOLOGY

In linguistics, semantics is contrasted to discourse and in philosophy, ontology is contrasted with epistemology. Discourse and epistemology can address how we develop, analyze, and talk about new knowledge. Because scholarship is concerned not only with the representation of exiting knowledge but also the development of new knowledge, we must incorporate structured ways of describing the acquisition of new knowledge. The exploration of claims about knowledge requires descriptions which are very different from those in formal semantic ontologies. Scholarship often starts by considering an unknown or “gap”,

Many systems of discourse structures have been developed for argumentation and explanation. Notable among the argumentation frameworks is AIF (Argumentation Interchange Format). This supports a variety of “argumentation schemes” [4] such as deduction, induction, and abduction. Yet, like most other discourse and argumentation frameworks, AIF is not closely coordinated with semantic structures. Claims must be evaluated by the quality of the evidence and by the relevance of the research to existing semantic models. Thus, the ontologies and argumentation models must have interwoven representations. Along the same lines, explanations are often based on the analysis of causal relationships which need to be incorporated into the semantic analyses.

4. UNIVERSALS AND PARTICULARS (SCIENCE AND HISTORY)

Rich semantics and structured applied epistemology [1] could initially be applied to individual scientific research reports. Indeed, this can be viewed as another step in the long-term trend of increasing structure in research reports [3]. Eventually, those research reports can be organized into highly-structured digital libraries.

The results from the research reports may entail changes to the reference ontologies and associated reference models. So we need to develop a full “structured applied epistemology”. In science, the criteria of quality and relevance for evaluating evidence are known as internal and external validity. Internal validity refers to whether the research was completed as intended. Thus, structured representations of the research designs are needed and typical problems with a research method must be checked. External validity refers to the implications of the results for reference models so frameworks for comparing and coordinating them are needed.

While scientific research often explores and extends models about classes (universals), science is also used to explain the causes of for specific events (instances). To handle this latter case, we need to model instances and to be able to link those instance models to the reference models. Beyond natural history, our representations should cover human and social histories; however, this introduces a new range of challenges. Although causation is generally accepted for scientific explanations, it much less clear for human and social histories. In turn, this means more ambiguity and controversy in the explanations given. While there should be value for richly structured descriptions of human and social histories, they need to include particularly robust support for argumentation and versioning of the analyses.

In addition to the histories discussed above, rich semantics and direct representation may also be applied to the organization and description of materials such historical newspapers, scholarly editions, and biographies.

5. IMPLEMENTATION AND IMPLICATIONS

This project should be considered a Grand Challenge and will take many years to accomplish. Nonetheless, its components are of considerable value in their own right. There would be several different types of repositories such as reference ontologies and models and instance models. The latter could be so massive as to need distinct sections such as the structured research reports we discussed in the previous section. In addition, there could also be less formal, but still highly-structured material such as work-in-progress and commentary.

The services associated with highly-structured repositories will be different from text-based repositories. For instance, text mining will be less important but text generation for summaries and tutoring will be more so. In addition, although traditional citations will not be eliminated, much of their utility may be replaced by direct linking of relevant sections of related documents.

Some of the initial steps toward highly-structured repositories may be relatively easy. We have previously urged the development of a comprehensive distributed collection of online historical materials. In addition, developing small collections of structured materials will be useful. However, scaling these will be difficult and require active community engagement. The Open Biomedical Ontology Foundry [2] is an example for how such community participation can be coordinated. A similar initiative might be introduced for developing the ontologies needed for societal models. Community involvement will also be required for policies and procedures in determining thresholds for “consensus” about results. In any event, despite the difficulties, the approach outlined here has the potential for greatly improving access and use of vast amounts of important information.

6. REFERENCES

[1] Allen, R.B., From Ontology to Epistemology, 2016, arXiv: 1610.07241v2

[2] Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., Goldberg, L.J., Eilbeck, K., Ireland, A., Mungall, C. J., Leontis, N., Rocca-Serra, P., Ruttenberg, A., Sansone, S.A., Scheuermann, R.H., Shah, N., Whetzel, P.L., and Lewis, S., The OBO Foundry: Coordinated Evolution of Ontologies to Support Biomedical Data Integration, Nature Biotechnology, 25, 1251–1255, 2007. DOI: 10.1038/nbt1346

[3] Swales, J.M., Genre Analysis: English in Academic and Research Settings, Cambridge, 2004.

[4] Walton, D., Reed, C., and Macagano, F., Argumentation Schemes, Cambridge U Press, 2008.