On the Mathematical Foundations of Semantic Interoperability and Integration*

Extended Abstract

Marco Schorlemmer

Institut d'Investigació en Intel·ligència Artificial
Consell Superior d'Investigacions Científiques
Bellaterra (Barcelona), Catalonia, Spain
marco@iiia.csic.es

1 Introduction

We discuss the mathematical foundations underlying the explicit use of semantic descriptions to facilitate information and systems integration. We do not attempt to provide a formal framework that is mature enough for modeling semantic interoperability and integration, nor do we attempt to fix the foundations of such a framework yet. Instead we want to stimulate the discussion around three main questions that need to be eventually addressed within any rigorous approach to semantic interoperability and integration.

How would a formal definition of semantic integration look like? There are many different uses of semantic integration terms, but little agreed upon terminology. In general, the community makes ambiguous uses of intuitive terms such as "equivalent", "equal", "sublcass of", "overlapping", "related to", etc., which seldom are defined in a rigorous way, i.e., with respect to a mathematical model. Hence, we lack of a theoretical framework upon which to define semantic integration terminology. But, is it possible to come up with such a theoretical framework? Is it actually desirable?

What do we require of a mathematical model for semantic integration? If we would to establish a mathematical framework in which to formalize and to model the intuitions concerning semantic integration, what should be its features and what should be its scope? Should it embrace the several approaches that have been explored so far, such as channels for the Barwise-Seligman theory of information, local-as-view approaches from database theory, alignment from ontology engineering, and blending from cognitive linguistics? It would be desirable that

* Breakout session held at the Dagstuhl Seminar on Semantic Interoperability and Integration on September 23, 2004. Participants: Joseph Goguen, Michael Grüninger, Peter Haase, Robert Kent, Werner Kuhn, Chris Menzel, Till Mossakowski, Marco Schorlemmer, Kean Huat Soon, Gerd Stumme.

Dagstuhl Seminar Proceedings 04391
Semantic Interoperability and Integration
http://drops.dagstuhl.de/opus/volltexte/2005/44

the intuitions underlying such mathematical model be understandable by knowledge engineers and practitioners in the field, but: Should it also be a framework stressing the pragmatic and practical side, or should it only define terminology in a rigorous way? Optimally it should be both.

Which would be the appropriate mathematical technique? The model-theoretic approach to semantics of first-order logic has always enjoyed a special status in knowledge representation and reasoning. But, is it the appropriate formalism to act as a mathematical framework for semantic interoperability and integration? Many alternative approaches to semantics have been advocated: possible-world semantics, property-theoretic semantics, situation semantics, etc., although current standardization efforts mainly stay within fragments of first-order model theory. Shall we stay within first-order because of its intuitiveness, or shall we draw from more abstract category-theoretic techniques such as institution theory?

2 The Role of Institution Theory

Institution theory arose in the 1980s as part of the effort of modularizing and parameterizing formal specifications of software systems [1]. An institution captures the essential aspects of logical systems underlying any specification theory and technology.

There seems to be consensus that institution theory might be the appropriate mathematical technique to describe semantic interoperability and integration at the general level, as it captures the key ingredients of every semantic interoperability scenario:

a notion of signature: Semantic interoperability is about meaningful exchange of symbolic items denoting classes, relations, attributes. Hence it is about signatures.

a notion of expressions over this signature: Semantic interoperability might involve the composition of signature symbols into more complex expressions such SQL queries or OWL expressions.

a notion of models that interpret the signature symbols: Semantic interoperability is eventually about semantics. Hence, it is about attaching meaning to signature symbols and expressions. We need a notion of model that provides meaning to signature symbols and expressions.

a notion of satisfaction between models and expressions: A way of relating expressions with models, and determining when an expression is true in a model, or a model satisfies the constraints determined by an expressions.

There exists an extensive literature on the topic, including many extensions of institutions that include notions of inference and proof. Important for the issues of semantic interoperability is the idea of an institution morphism [2], that is, a way of describing transformations between institutions that preserve their essential structure (a way of transforming signatures, expressions and models, respecting the satisfaction relationship between models and signatures).

3 A Mathematical Model For Semantic Integration

Institution theory seems to be the right mathematical tool for providing a precise definition of the semantic heterogeneity problem, and programming language designers have used results about institutions in designing advanced module systems. But using the *full-blown* theory is not at the right level of abstraction to make it suitable for knowledge engineers and practitioners:

1. *It is too abstract:* This is both the strength and the weakness of the theory. Its abstractness makes it suitable to address the core aspects of semantic heterogeneity without committing to any particular formalism or language. But this makes it hard as a tool for tackling concrete semantic interoperability problems.
2. *It requires too much technical knowledge:* Institution theory is based on category theory [3], which is a highly abstract branch of mathematics that spun off from topology based on the observation that many concepts across different fields of mathematics could be unified by focusing on structure-preserving transformations between mathematical structures. Mastering the concepts of category theory and institution theory requires significant amounts of effort and dedication.
3. *It provides no direct link to practical problems:* Institution theory is a powerful theoretical framework, but does not provide immediate insight into the practical problems faced by knowledge engineers.

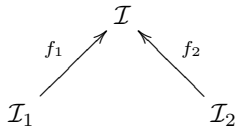
What we need is a theoretical framework that fixes the fundamental ideas of the institutional framework hiding the category-theoretical machinery. Category theory and thus institution theory is unfamiliar even to most mathematicians and logicians. No one developing semantic integration methods and tools should ever be faced with institution theory directly, but should be able to work with rigorous semantic integration theory that is built upon it.

In addition, more efforts trying to bridge the gap between the theoretical framework and its application to concrete problems are required, such as systems like [4] with nice GUIs with absolutely no institutions anywhere in sight, but certainly doing a lot of work under the hood.

4 Towards a Formal Definition of Semantic Integration

On the abstract level, in the institution-theoretical framework, a tempting definition could be:

Definition 1. *Two systems are semantically integrated if, and only if, there is a co-relation between their underlying institutions:*



Almost every semantic interoperability scenario will eventually be a special case of co-relation between institutions. Sometimes, though, there are additional objects in the co-relation diagram, for shared material, so that one should take a pushout, or more generally, a colimit, to get a smallest solution; but sometimes, other notions of optimal semantic integration may be more appropriate than the universal solution idea of limits and colimits.

Thus, how does the above definition shed light into the semantic heterogeneity problem in the first place? What are these institutions? How can they help the knowledge engineer and practitioner to establish appropriate co-relations between them?

These are all questions that need to be addressed in any formal framework that provides a definition of semantic integration. For the knowledge engineer or practitioner, such a definition should mark the path for developing semantic integration methods and tools which are sound with respect to the its theoretical foundations. The institution-theoretical framework is still far away from this goal.

5 Future Work

As "pragmatic theoreticians" concerned with the formal foundations of our discipline, but also aware of the practical needs of knowledge engineers and practitioners, we need to focus on the development of an appropriate mathematical framework for semantic interoperability and integration that draws from the vast foundational work on institution theory but presents it in a form suitable to tackle practical needs arising from the semantic heterogeneity problem.

The importance of the problem and the lack of rigorous theoretical foundations reveal the urgency to work towards this goal, by:

1. Performing **case studies of concrete semantic integration scenarios** within the institution-theoretic framework, i.e., defining particular institutions and institution morphisms, and establishing the necessary co-relations between institutions to reveal "dirty details" popping up when instatiating the abstract framework on concrete example. Some possible case studies could be:
   - co-relating institutions for theories based on time intervals and time points
   - co-relating institutions of RDF(S), OWL, KIF, ...
   - ...

   On would benefit greatly from practical examples. Consequently, it would be useful if the example systems came from a readily graspable application domain (rather than from category theory itself). Maybe two simple services (operations) that interoperate?
2. Based on the results and experiences drawn from the above case studies, **fixing the fundamental ideas** based on familiar mathematical concepts

(maybe drawn from intuitive set theory) and avoiding unfamiliar category-theoretic terminology. On could, for instance, try to define institutions without any category theory, although the mathematics may result considerably more complex looking than it would be if categories are used.

3. Slowly developing **a body of theory of semantic integration** that includes practitioners into the loop, in order to make it both mathematically rigorous and also useful for the knowledge engineer or practitioner for developing methods and building tools to support the task of semantic interoperability and integration.

References

1. J. Goguen and R. Burstall. Institutions: Abstract model theory for specification and programming. *Journal of the ACM*, 39(1):95–146, 1992.
2. J. Goguen and G. Roşu. Institution morphisms. *Formal Aspects of Computing*, 13:204–307, 2002.
3. S. Mac Lane. *Categories for the Working Mathematician*. Springer, second edition, 1998.
4. T. Mossakowski. *Heteregeneous Specification and the Heterogeneous Tool Set*. Habilitation thesis, Universität Bremen, 2004.