Semantic Interoperability and Integration

Dagstuhl Seminar 04391, September 19-24, 2004

Executive Summary

Y. Kalfoglou[1], M. Schorlemmer[2], M. Uschold[3], A. Sheth[4], and S. Staab[5]

[1] University of Southampton, UK
y.kalfoglou@ecs.soton.ac.uk

[2] Consejo Superior de Investigaciones Científicas, Spain
marco@iiia.csic.es

[3] The Boeing Company, USA
michael.f.uschold@boeing.com

[4] University of Georgia, USA
amit@cs.uga.edu

[5] Universität Koblenz-Landau, Germany
staab@uni-koblenz.de

1 Motivation

Semantic interoperability and integration is concerned with the use of explicit semantic descriptions to facilitate information and systems integration. Due to the widespread importance of integration, many disparate communities have tackled this problem. They have developed a wide variety of overlapping but complementary technologies and approaches. The seminar had the following objectives:

1. To stimulate collaboration between diverse communities bound by common objectives in the area of semantic interoperability and integration;
2. To lay the foundation for a framework and a theory for understanding and classifying technologies for semantic interoperability and integration;
3. To set the research agenda for this area with the long-term aim of building a "research pipeline" for creating and disseminating results in industry.

1.1 To Stimulate Collaboration

We shared experiences and raised awareness of goals and relevant technologies among relatively independent communities such as: database integration, digital libraries, information retrieval, ontologies, knowledge representation, category theory, semantic web (including services), agents, applications, interoperability and middleware (e.g., as specified by the OMG). For example, the database community described their "lessons learned" from semantic data models, schema and information system integration and heterogeneous information integration; the ontology community demonstrated the difficulties of ontology

mapping and merging; theorists pinpointed which theories are good candidates for understanding the problem and outlined the trade-off between expressiveness and computability; application developers explained what assumptions are made when building integrated systems; industry practitioners highlighted the real-world challenges and applications that provided realistic scenarios leading to deeper insights that academic prototypes may not encounter, discussed semantic technologies that have had opportunity to mature in commercial settings, and provided early examples of successes.

One of the seminar objectives was the development of a common terminology to improve communication and foster tighter collaboration between members from different scientific areas and background.

1.2 A Framework and Theory

We sifted through and analyzed the wide variety of approaches, technologies and experiences shared by workshop participants. This will lay the foundation for the future development of a comprehensive theory and reference framework for understanding and classifying the many theoretical and practical approaches to semantic interoperability and integration. The framework needs to be sufficiently abstract to be broadly agreed and pragmatic enough to be computationally applied and practically useful. The framework will be described informally and represented in order to allow certain degrees of automation.

1.3 Setting a Lasting Research Agenda

In the context of the above said reference framework, we planned to formulate an agenda for promoting collaboration and information exchange portal in this field. We believe that progress on semantic interoperability and integration will take a truly multi-disciplinary approach. Our objective is to foster collaboration involving academic research and lively academia-industry exchanges. This will be facilitated by such tools as (a) a set of matrices for measuring the effectiveness of various approaches (reminiscent of TREC and Reuter benchmarks), and (b) guidelines for researchers and practitioners who wish to apply the tools for integration purposes.

2 The Seminar Structure

A total of 39 researchers from academia, government and industry attended the seminar, 33% of which were young researchers aged under 35. Of those participants, 24 were affiliated with institutions from EU member states, 14 from the USA, and 1 from a Swiss institution. Of the EU participants, 12 came from Germany, 3 from the UK, 3 from Greece, and 1 each from France, Italy, Belgium, Spain and The Netherlands (see Appendix **??**).

As a consequence of the large audience, it was both impossible and undesirable to schedule separate talks for everyone. That would also be contrary to the spirit of the Dagstuhl seminars.

To keep things lively and to provide an opportunity for everyone to participate in a meaningful way, we included a variety of formats for cross-community interaction including feature presentations, shorter talks, panel discussions and breakout sessions—as well as a whole afternoon off on Wednesday, which we devoted to hiking in the area, or to a sightseeing trip to Trier and the Mosel valley.

Feature Presentations: There were a small number of in depth feature presentations from prominent members of the community. These were surveys of a particular area, or a close look at a particular technique, approach or project—or a combination thereof. The goal was to present update material to the multi-disciplinary audience.

Short Talks: There were a larger number of conference-style talks by various individuals.

Breakout Sessions: Small group sessions were sprinkled throughout the week to facilitate in depth discussions on various special topics to address the seminar objectives. The topics were decided on during the week.

Panel Discussions: Important and/or controversial topics were the subjects for various lively panel discussions.

In the early part of the week, the major emphasis was in getting to know each other and the broad range of challenges and results in the field. In the middle and later parts of the week, there were more interactive sessions, fewer talks and an opportunity to delve into various topics at length and in depth. At the final day of the seminar, we examined what we learned and identified important steps for advancing the state of the art and improving the state of the practice.

3 Highlights of the Week

Feature talks, short talks, and panel discussions were structured roughly around four main themes: *Mapping and translation*, *industrial experiences*, *theoretical foundations*, and *standards and benchmarks*. Different days were devoted to different themes.

3.1 Mapping and Translation

On Monday, Natasha Noy and Heiner Stuckenschmidt opened with a feature talk on *Ontologies: Mapping, Translation, Merging*. They distinguished different types of heterogeneity between ontologies, discussed various mapping representations, classified methods for discovering mappings both between ontology classes and instances, and talked about various tasks where mappings are used.

These issues were further addressed in short talks given by Martin Doerr on *The CIDCOM CRM: An Ontological Approach to Schema Heterogeneity*, by Fausto Giunchiglia on *Semantic Matching* and by Gerd Stumme on *Ontology Merging with Formal Concept Analysis*.

The first day concluded with a panel moderated by Amit Sheth on *The Role of the Web Services in Semantic Interoperability and Integration*. Panelists were Andreas Abecker, Mark Bernstein, Vassilis Christophides, and Tim Finin. They discussed the importance of application interoperability, despite of the fact that the semantics of programs wrapped by Web Services is much harder to understand than the semantics of data. Beyond capturing the domain semantics of data object manipulated and exchanged by Web Services, one had to think about functional semantics, execution semantics and Quality of Service issues It was pointed out that although total standardization will never happen, standardization is a preferred solution to run-time mediation when possible. Web services provide a useful basis for semantic interoperability and integration, as well as introducing new problems and research challenges. Additionally, semantic interoperability and integration tools will play a central role as services on the Semantic Web themselves.

3.2 Industrial Experiences

Naveen Ashish and Amit Sheth initiated the Tuesday session reporting on *Industrial Experiences*. Ashish provided an overview of several ongoing NASA endeavors based on concepts, systems, and technology from the Semantic Web arena, such as completed R&D efforts for several applications ranging from collaborative systems to aerospace information management to enterprise search to scientific information gathering and discovery systems at NASA. Sheth presented a semantic technology with origins in academic research that has been commercialized and deployed in supporting significant industrial and government applications. He outlined core functional capabilities that have proved to be important, and made a series of observations based on real-world applications of potential interest to the Semantic Web research community. These involved the topics of ontology quality, ontology language expressiveness, ontology size and freshness, automatic semantic metadata extraction, content heterogeneity, and performance needs of semantic computing capabilities.

These talks were followed by an open discussion on *Use Cases and Requirements for Semantic Interoperability and Integration* led by Steffen Staab. Hans Chalupsky, Robert Meersman, and Werner Kuhn identified several interoperability scenarios ranging from geospatial applications and helicopter mission planning to VAT regulations and the tourism industry. It was stressed that domain ontologies need to be built and vetted by domain experts and scientists, as those built by *computer* scientists were usually rejected.

The session was complemented with a short talk by Vassilis Christophides on *Integrating XML Data Sources using RDF/S Schemas: The ICS-FORTH Semantic Web Middleware (SWIM)*.

The remainder of the afternoon was devoted to several breakout sessions, formed around specific topics: *Case Studies of Semantic Interoperability and Integration* led by Vasilis Vassalos, *Mapping Typology* led by Martin Doerr, and *Mapping Notations and Languages* led by Mike Uschold.

During the evening we had an open and extremely lively discussion over cheese and wine led by Amit Sheth on *Lessons learned and future trends* for semantic interoperability and integration. We learned about key results published by the database and information systems community relevant to ontology matching, types of mappings and the use of ontologies and description logic in databases/information systems research. Although much of this work was done 10-15 years ago, it is relevant to contemporary work on semantic interoperability and integration. An interesting distinction that came out during the discussion was the recognition that while most work in the DB community was under a closed-world assumption, current work in ontology-based semantic integration is based on an open world assumption.

3.3 Theoretical Foundations

On Wednesday, Michael Grüninger and Chris Menzel gave a feature talk on *PSL: An Industrially Motivated and Rigorously Formal Approach to Semantic Integration*. They addressed the problems of generating semantic mappings between ontologies, determining that they are correct, and providing a vehicle for executing the mappings, thus translating terms from one ontology into another. The techniques for semantic integration proposed in their talk exploited the model-theoretic structures of the underlying ontologies to support the semi-automatic generation of mappings and the automatic comparison of such mappings.

Additional work on the topic of theoretical foundation for semantic interoperability and integration was reported in short talks by Robert Kent on *Semantic Information in the Information Flow Framework* and by Till Mossakowski on *Heterogeneous Specification and the Heterogeneous Toolset*.

3.4 Standards and Benchmarks

The Thursday session started with a feature talk by Peter Denno and Jérôme Euzenat on *The Role of Standards and Benchmarks in Semantic Interoperability and Integration*. They surveyed the purpose of standards, standards' bodies, their various procedures and culture. They asked how semantic interoperability might benefit from standardization, considering how standardization of ontologies may differ from standardization of anything else. They also presented two efforts toward evaluating ontology alignment programs that have been run in 2004 and drew lessons from them.

During the remaining of the morning session Vasilis Vassalos gave a short talk on *MiniCount: Answering Aggragate Queries Using Views* and Michael Sintek one on *Using TRIPLE Views for Semantic Interoperability and Integration*.

In the afternoon, further breakout sessions were formed. Tuesday's *Mapping Typology* session continued with a shift in emphasis towards *Mapping Tools*. This was led by Dagobert Soergel. New topics also emerged: sessions were held on *Social Aspects in Semantic Interoperability and Integration* led by Klemens Böhm, *Infrastructure and Architecture* led by Mike Uschold and *Theoretical Foundations for Semantic Interoperability and Integration* led by Marco Schorlemmer,

The day ended with a lively panel led by Yannis Kalfoglou on *Heavy-weight versus light-weight approaches: Bridging the Gap Between Theory and Praxis.* Panelists were Dave Robertson, Joseph Goguen, and Karl Aberer. The discussion revolved around the application of formal methods and the importance of pragmatic issues. It was pointed out that field research suggest that scientists want things as lightweight as possible, although heavyweight approaches are needed when there are enormous amounts of data, processes are well understood and there is much stability. Heavy-weight approaches would comprise centralized, model-centric, human controlled and standard-based approaches, while lightweight ones would be distributed, inference-centered, automated, self-organized and mediated.

4 Breakout Sessions

The concluding session on Friday morning was devoted mainly to report back on the discussion and outcomes of the various breakout sessions and in establishing future actions in the field of semantic interoperability and integration.

4.1 Social Aspects

A small group of participants decamped to discuss and debate the role of social aspects and how these could affect an interoperability and integration strategy. One of the issues discussed, was the origin of ontologies, a major vehicle in achieving interoperability. It was pointed out that ontology generation should be done by community members rather than a handful of skillful engineers. That raised the question of how to increase human involvement in the process: it was argued that socially-inspired computing is different from social engineering, a norm in everyday practice at organizations. More community involvement could be achieved using peer to peer technology. It was noted that large peer to peer communities share and aggregate personal data. These could be used to help with interoperability and integration but it is not clear how. There is much AI literature on dialog, perhaps this could apply in the peer to peer context? There was some discussion about privacy issues as well as conflict/disagreement resolution.

4.2 Use Cases and Requirements

This session was motivated by a desire to identify target 'customers' of semantic interoperability and integration technologies. What kinds of requirements are there? Can we identify any specific use cases and scenarios? By adding a demand-pull aspect of research, early success stories are more likely. The session unfolded as an exercise of cross-community communication about problems in semantic interoperability and integration. Each of the participants proposed a different problem, explained it in enough detail so that others could begin to work on it and fit it into their mental models. The problems ranged from the sufficiency of

maintenance of inference for semantic integration, to the problem of automated structuring of queries based on keywords with help of domain ontologies, to sharing of inferences among reasoning engines based on different logical systems. A question was raised about the fact that there seems to be a dearth of specific use cases that have been identified.

4.3 Mapping Typology and Tools

The objectives of this breakout session were to establish frameworks for classifying tools in terms of application-relevant categories and for defining case-specific requirements for mappings, and to identify objectives for further research. This group worked through the various dimensions of mapping requirements and how to resolve terminological clarifications. Specific challenges were identified, such as instance relation matching, controlled information loss, interaction of human input and automated solutions, and tools refining existing mappings.

In a continuation of this session, the issue of tool support was addressed: *What tools are out there? What functionality is required? What is being mapped? What types of mapping are being discovered? In what format are they represented? How much computer assistance is offered? What are the right measures of success?*

4.4 Mapping Notations and Languages

The goal of this session was to better understand how we can represent the mappings that are critical to achieving semantic interoperability and integration. What actual notations or languages are being used today? To what extent are they meeting the requirements? Is there any tool support? What are key important problems that the research community needs to address?

First, various aspects of mappings were identified. These included: what is being mapped, declarative vs procedural mappings, translation engines for executing mappings, composing mappings and the role of inference. Another key issue is the expressiveness of the mapping language, as compared to the expressiveness of the ontology language of the mapped ontologies. The purpose for mapping also varies: e.g., for querying, vs. data translation.

Next, we discussed the mapping languages that are currently available for use. The Semantic Web ontology language, OWL was specifically designed to support ontology mapping, however the constructs provided are very limited. Being purely declarative, it does not support various procedures that are needed to perform mappings for data translation (e.g. arithmetic and string operations). There are systems available to do this today: for example, N3 rules in conjunction with CWM; datalog, F-Logic and DLR. Currently, the W3C is developing SWRules which will be able to do some of these things.

Finally, we discussed open research problems. Issues identified were: ontology language heterogeneity, expanding the expressiveness of mapping relations and viewing mappings as first-class citizens.

4.5 Theoretical Foundations

Three main questions were addressed during this breakout sessions: What would a formal definition of semantic integration look like? What do we require from a mathematical model for semantic integration? What mathematical techniques are appropriate?

Initial consensus emerged that the theory of institutions (an abstract description of logical systems) seems to be the right mathematical tool for providing a precise definition of the problem. A caveat is that this approach is unlikely to be appropriate for the practitioner. Thus a theoretical framework that fixes the fundamental ideas of institution theory needs to hide the category-theoretical machinery.

Hence it was suggested to do case studies for applying the institutional framework to reveal "dirty details", by relating theories based on time intervals and time points, or by establishing institution morphisms between RDF(S), OWL, KIF. This would fix the fundamental ideas based on these case studies and help developing a theory of semantic integration that is both mathematically rigorous and useful for the knowledge engineer.

4.6 Infrastructure and Architectures

This session was aimed at determining what kinds of infrastructural and architectural support may be needed for semantic interoperability and integration. There should be support for multiple or hybrid architectures (e.g., point to point, vs. mediated). Also, in a Semantic Web context, it will likely be necessary to support registering different ontologies, mappings between them, translation engines that can execute mappings etc. This would allow dynamic semantic interoperability and integration to take place.

Fundamental to dynamic interoperability is the need for self-describing agents and services. Importantly, such descriptions will be required at multiple levels of abstraction. Protocols are a key infrastructural element. Lacking a single standard, agents will have to negotiate which protocol to use, and be competent in more than one. Agents also need to self-describe their goals and their ontology(s). Mappings between different ontologies also need to be made available, although they are not agent-specific. Finally, there needs to be architectural/infrastructural support for plan/goal decomposition.

5 Conclusion

We set out to achieve three main objectives in the area of semantic integration and interoperability: 1) to stimulate collaboration, 2) to lay the groundwork for a future comprehensive framework for understanding the field and 3) to set a research agenda.

We successfully brought together experts from industry, academia and government representing historically separate communities including: database integration, category theory, standards, digital libraries, ontologies, knowledge representation, and the semantic Web. The varied program including invited talks,

shorter talks, panels and breakout sessions provided the context for much collaboration. We achieved this objective. Various outputs for the workshop are captured on the Web at: http://www.dagstuhl.de/04391/Materials/

Much of the week was spent wrestling with differences in terminology, identifying a range of relevant broad issues, puzzling over more specific and often subtle distinctions that arise in different sub-areas. The topics for the breakout sessions, form a good starting point for the eventual development of a more comprehensive framework that will provide a background for understanding and comparing different techniques, tools and applications that are developed in coming years.

Finally, we considered a range of issues that need to be included on agenda for future research. The problem of semantic interoperability and integration is hard and it is not clear how it could be solved in the near future. For instance, we don't know how to formally specify the problem yet, let alone solving it. Hence it is necessary to work on two fronts: theoreticians need to make their case for the appropriate foundations upon which semantic interoperability and integration can be formalized, while practitioners and users need to expose their local semantics for the benefit of knowledge sharing.

The Semantic Web provides a playground for experimentation, but it also introduces new problems. There are a lot of challenging infrastructure issues still to be addressed and standardization efforts are still at their infancy. Furthermore there is a lack of lengthy experiences and large-scale scenarios to evaluate the scalability of current methods and techniques.

The discussions and outcomes of talks, breakout sessions and panels during the seminar have highlighted these issues and helped to put together efforts which were previously conducted separately in different communities. More discussion and cross-disciplinary collaboration is needed, but the first steps in converging and reaching a consensus might already be well under way.

A Participants

Andreas Abecker (Forschungszentrum Informatik, Germany)
Karl Aberer (École Polytechnique Fédérale de Lausanne, Switzerland)
Naveen Ashish (NASA Ames Research Center, USA)
Mark Burstein (BBN Technologies, USA)
Christoph Bussler (National University of Ireland, Ireland)
Klemens Böhm (Universität Magdeburg, Germany)
Hans Chalupsky (University of Southern California, USA)
Vassilis Christophides (Foundation for Research and Technology, Greece)
Ernesto Compatangelo (University of Aberdeen, UK)
Peter Denno (National Institute of Standards and Technology, USA)
Martin Doerr (Foundation for Research and Technology, Greece)
Marc Ehrig (Universität Karlsruhe, Germany)
Jérôme Euzenat (Institut National de Recherche en Informatique et en Automatique, France)
Timothy Finin (University of Maryland, USA)

Fred Fonseca (Penn State University, USA)
Fausto Giunchiglia (Università di Trento, Italy)
Joseph Goguen (University of California, USA)
Michael Gruninger (National Institute of Standards and Technology, USA)
Peter Haase (Universität Karlsruhe, Germany)
Yannis Kalfoglou (University of Southampton, UK)
Robert Kent (The Ontology Consortium, USA)
Werner Kuhn (Westfälische Wilhelms-Universität Münster, Germany)
Robert Meersman (Vrije Universiteit Brussel, Belgium)
Chris Menzel (Texas A&M University, USA)
Till Mossakowski (Universität Bremen, Germany)
Erich Neuhold (Fraunhofer-Gesellschaft, Germany)
Natasha Noy (Stanford University, USA)
Dave Robertson (University of Edinburgh, UK)
Marco Schorlemmer (Consejo Superior de Investigaciones Científicas, Spain)
Amit Sheth (University of Georgia & Semagix Inc., USA)
Michael Sintek (Deutsches Forschungszentrum für Künstliche Intelligenz, Germany)
Dagobert Soergel (University of Maryland, USA)
Kean Huat Soon (Westfälische Wilhelms-Universität Münster, Germany)
Steffen Staab (Universität Koblenz-Landau, Germany)
Heiner Stuckenschmidt (Vrije Universiteit Amsterdam, The Netherlands)
Rudi Studer (Universität Karlsruhe, Germany)
Gerd Stumme (Univesität Kassel, Germany)
Mike Uschold (The Boeing Company, USA)
Vasilis Vassalos (Athens University on Economics and Business, Greece)