Semantic Interoperability in Global Information Systems

A brief introduction to the research area and the special section

Aris M. Ouksel and Amit Sheth

Internet, Web and distributed computing infrastructures continue to gain in popularity as a means of communication for organizations, groups and individuals alike. In such an environment, characterized by large distributed, autonomous, diverse, and dynamic information sources, access to relevant and accurate information is becoming increasingly complex. This complexity is exacerbated by the evolving system, semantic and structural heterogeneity of these potentially global, cross-disciplinary, multicultural and rich-media technologies. Clearly, solutions to these challenges require addressing directly a variety of interoperability issues.

One can define several forms of interoperability in information systems. Figure 1 shows one of several classifications that presents the interoperability types based on various forms of perspective on heterogeneity in information systems (cf: Sheth 98). Focusing on the crucial dimension of heterogeneity and corresponding solutions leads us to discuss different levels of interoperability—*system, syntax, structure*, and *semantic*. In this classification, we consider differences in machine-readable aspects of data representation, also referred to as formatting, to be relevant to syntactic heterogeneity. We consider representational heterogeneity that involves data modeling constructs to be relevant to structural interoperability. Schematic heterogeneity that particularly appears in structured databases is also an aspect of structural heterogeneity. While significant progress has been achieved in system, syntactic, and structural/schematic interoperability, comprehensive solutions to semantic interoperability remain elusive (Ouksel 92, Ouksel and Iqbal 99). Yet, several trends and advances in software technologies are continuing to bring focus to semantic issues and semantic interoperability. This is the topic of this special section.

A more general framework for interoperability is illustrated in Figure 2. In this framework (Ouksel 99), it is recognized nuanced approaches to semantics and semantic interoperability are necessary. It is argued that current theories are insufficient to account for a

Information Heterogeneity Semantic Heterogeneity Structural, Representational/Schematic Heterog. Syntactic, Format Heterogeneity
System Heterogeneity **Information System Heterogeneity** Digital Media Repository Management Systems Database Management Systems (heterogeneity of DBMSs, data models, system capabilities such as concurrency control and recovery) **Platform Heterogeneity** Operating Systems (het. of file system, naming, file types, operation, transaction support, IPC) Hardware/System (heterogeneity of instruction set, data representation/coding)

Semantic Interoperability
Structural Interoperability
Syntactic Interoperability
System Interoperability

Figure 1. Heterogeneity in information systems and corresponding interoperability concerns.

variety of misinterpretations in a realistic social environment, which modern sophisticated applications demand. These theories are inadequate for supporting the dynamic integration of autonomous and heterogeneous information sources with possibly evolving and incompatible internal semantics, and ignore several other aspects of heterogeneity, particularly pragmatics. The framework posits semantics as a matter of continuous negotiation and evolution in an environment of uncertain and incomplete information, which preserves the autonomy of the information sources, and yet allows collaboration and cooperation in the presence of conflicts.

SOCIAL WORLD -- beliefs, expectations, commitments, contracts, law, culture, ...
PRAGMATICS -- intentions, communication, conversations, negotiations, ...
SEMANTICS -- meanings, propositions, validity, truth, signification, denotations, ...
SYNTACTICS -- formal structure, language, logic, data, records, deduction, software, file, ...

Figure 2. Open Systems Framework for Social Interaction

Another approach to support a more general notion of semantics is to relate the content and representation of information resources to entities and concepts in the real world (Beech 1997; Meersman 1997; Sheth 1997). That is, the limited forms of operational and axiomatic semantics of a particular representational or language framework are not sufficient (see Paepcke et al. 1998 for a relevant discussion on syntax and some types of semantics, also see Lee et al 1996 for a logic and knowledge based perspective). Semantic interoperability will then support high-level (hence easier to use), context-sensitive information requests over heterogeneous information resources, hiding system, syntax, and structural heterogeneity. In essence, we need an approach that reduces the problem of knowing the contents and structure of many information sources to the problem of knowing the contents of easily-understood, domain-specific ontologies, which a user familiar with the domain is likely to know or understand easily. During the 1980s when we were working towards integrating multiple databases and their schemas, our concern was to identify objects that were represented differently but were related conceptually-- that is we were interested in "So for Schematically, yet So Near Semantically" (Sheth and Kashyap 1993) . With the massive information overload in the global information infrastructure when a query may return thousand of results we a user may ill afford to go through, our emphasis seems to have shifted "So near Syntactically/Schematically, yet so far Semantically".

Foundational research leading to building the new generation of global information systems that support semantic interoperability has been carried out in several umbrella projects and initiatives, including Knowledge Sharing Effort (http://www-ksl.stanford.edu */knowledge-sharing*), Intelligent Integrationof Information (*http://mole.dc.isx.com/I3*), and the Digital Library Initiative (http://www.cise.nsf.gov */iis /dli_home.html*). Increasing standardization at different levels of information systems architecture for corresponding type of interoperability also plays an important role. Some of the examples are as follows.

- *System*: IIOP for interactions between distributed objects and components, KQML for interaction between agents;
- *Syntactic*: XML for all forms of Web-accessible data;
- *Structural*: RDF for general purpose description of information sources, various object models for web-based information exchange (Manola 1998), MPEG-4 for structural or object-level description video, MHEF-5 for multimedia and hypermedia, KIF for knowledge representation, OKBC for distributed knowledge bases;
- *Semantic*: MPEG-7 (still in progress) with likely support for limited forms of semantics with identification of context, objective requirements, and applications.

Sophisticated approaches to semantic interoperability are motivated by several trends in software technologies and organizationally complex information infrastructures. These include:

- ease of accessing and publishing a broad variety of data and data sources, with the corresponding challenge in heterogeneity and information overload from using simpler (such as keyword based) access techniques
- progress in techniques to model, capture, represent and reason about semantics; graduate progress in attention from data to information, and increasingly knowledge
- challenges in dealing with non-traditional (esp. visual) data that cannot be easily handled with well known IR and traditional database techniques

- attention to the issue of interoperability in various domains and research areas (e.g., bibliographic data, digital libraries, geographic and environmental data, space and astronomy data, etc.) and the improved technological ability
- support to the evolving concepts of virtual organizations and adhocracies -- and concomitant requirement for flexible semantic interoperability to interpret the available information in light of new market contingencies and the variety of intra- and cross-disciplinary forms of collaboration scientific or otherwise.

We now focus our attention on a discussion of possible enablers of semantic interoperability. In particular, we identify four enablers and capabilities:

Terminology (and language) transparency: This will allow a user to choose an ontology of his or her choice (e.g., one based on LCC for querying bibliographic data or FGDC for geospatial data), while allowing the information source to subscribe to a related but different ontology (e.g., an ontology based on DDC or UDK, respectively. The latter recognizes some overlap between geospatial data sets and environmental data sets, and their respective modeling).

Context-sensitive information processing: The information system will recognize or understand the context of an information need and use it to limit information overload, both by formulating more precise queries used for searching information sources and by filtering and transforming the information before presenting it to the user.

Rules of interaction mechanisms: This is not a standardization of semantics as in ontologies. Rather, these mechanisms formally specify the format of messages and the data types on communicated semantic and pragmatic information without any infraction on the substance being communicated, and the exchange protocols. We referred to these rules of communication as Semantic Cooperation Protocols (SCPs) (Ouksel, 1992). These rules provide means for the interacting parties to reach agreements on norms, responsibilities and commitments.

Semantic correlation: This will allow the representation of semantically related information regardless of distribution and heterogeneity (including various forms of media) by the user or the third party, and their use for obtaining all forms of relevant information anywhere.

to develop more challenging applications (e.g., digital earth, digital human) involving wider variety of users and perspectives over shared information resources.

Three key components of a possible solution are metadata (especially domain-specific and content-based metadata), contexts (Ouksel and Naiman, 1994), and ontologies (Kashyap and Sheth 1998). We briefly discuss their role in developing semantic interoperability solutions.

Ontologies and terminology transparency

An ontology can be defined as a specific vocabulary and relationships used to describe certain aspects of reality, and a set of explicit assumptions regarding the intended meaning of the vocabulary of words (Gruber 1991; Guarino 1998). Among various other classification schemes (Ouksel 1992, Naiman and Ouksel 1995) and structures, including keywords, thesauri, and taxonomies, ontologies are often viewed as allowing more complete and precise domain models (Huhns and Singh 1997). Support and use of multiple, independently-developed ontologies is important for developing scalable information systems with multiple information producers and consumers (e.g., Arens et al. 1996; Dao and Perry 1996; Genesereth and King 1995; Kashyap and Sheth 1998; Khang and McLeod 1998 for need and use of multiple ontologies; Ouksel and Iqbal 1999). One challenging issue in supporting semantic interoperability is how to allow both users and providers to subscribe to existing ontologies of their choice or create a new one (Kashyap and Sheth 1998). Processing an information request represented in terms of one ontology in an environment with information resources that subscribe to different (but related and relevant) ontologies may involve using inter-ontological relationships, such as synonym, hypernym, homonym, and other possibly domain-specific relationships. This work also requires understanding of and containing loss of information in multi-ontology query processing (Mena et al. 1998). One early example of research along these lines is the OBSERVER (sub)system (*http://siul02.si.ehu.es/~jirgbdat /OBSERVER*), which is a component of the InfoQuilt system (http://lsdis.cs.uga.edu/infoquilt).

From a theoretical point of view, it is important to note an important caveat about the

sufficiency of ontologies to resolve semantic conflicts. We contended in (Ouksel and Iqbal 1999) that while ontologies are useful in semantic reconciliation and are indeed necessary for practical and performance considerations, they do not guarantee in and of themselves correct classification of semantic conflicts, nor do they provide the capability to handle evolving semantics or a mechanism to support a dynamic reconciliation process. In constructing ontologies, rigid assumptions are generally made about commensurability of knowledge and the semantics and pragmatics of the interacting agents to achieve the understandable goal of precision and disambiguation. In (Ouksel and Iqbal 1999) we pointed out the limitations of this approach in dealing with semantics, even in a specific domain. We illustrated the deficiencies of ontologies through a conceptual analysis of several prominent examples used in heterogeneous database systems and in natural language processing. This analysis resulted in important outcomes. It allowed us to synthesize some essential features of semantic reconciliation. Semantic reconciliation is a non-monotonic query-dependent process that requires flexible interpretation of query context, and a mechanism to coordinate knowledge elicitation while constructing the query context. These features underpinned the design of the SCOPES architecture (Ouksel and Naiman 1994, Ouksel 1999), and are also recognized in (Scott McKay 1999, this issue). Clearly, in our view, work on ontologies presents enormous challenges and current assumption require further scrutiny.

Context

In characterizing the similarity between objects based on the semantics associated with them we have to consider the real-world semantics (RWS) of an object. It is not possible to completely define what an object denotes or means in the model world. We propose the *context* of an object as the primary vehicle to capture the RWS of the object. Understanding of the context of the information request can help the system to distinguish between whether the term *cricket* refers to an insect or a sports game.

Adapting from research in AI and Knowledge-Based systems (e.g., Shoham 1991), linguistics and other fields, modeling and representing context can lead to several benefits in dealing with information overload in a global information infrastructure/ systems (see Kashyap and Sheth 1998 for more details):

- *Economy of representation*: In a manner akin to database views, contexts can act as a focusing mechanism when accessing the component databases or information sources on the global information systems.
- *Economy of reasoning*: Instead of reasoning with the information present in the database as a whole, reasoning can be performed with the context associated with an information source.
- *Managing inconsistent information*: In the global information systems, where information sources are designed and developed independently, it is not uncommon to have information in one source be inconsistent with information in another. As long as information is consistent within the context of the query of the user, inconsistency in information from different databases may be allowed.
- *Flexible semantics*: An important consequence of associating abstractions or mappings with context is that the same two objects can be related to each other differently in two different contexts. Two objects might be semantically closer to each other in one context as compared to the other.

There are several proposals for representing context. We believe that an effective approach needs to bring together metadata, user profiles, information modeling abstractions, and ontologies, as well as to allow their dynamic construction to model application domain and user needs. Besides their modeling and representation, a key challenges includes the ability to reason about or compare contexts (e.g., Kashyap and Sheth 1996; Lee et al. 1996; Ouksel and Naiman 1994). While there are many representations and associated reasoning techniques, practical application of context in GII is expected to be a key research challenge for achieving semantic interoperability in information systems.

Information co-relations

One of the key applications of semantics in global information systems is to represent or specify information requests and semantic level information

co-relations regardless of the media (and other heterogeneity) and locations of information sources. These can involve queries over heterogeneous media assets represented at a higher level of abstraction in media-independent manner, using metadata and ontologies.

Two approaches to *representing* information correlations between independently managed networked resources are Metadata Reference Links (MREFs; Shah and Sheth 1998) and Distributed Active Relationships (DARs; Daniel et al. 1998). They provide an initial step in specifying information correlation between heterogeneous digital media. Specifically, MREFs allow subscription to one or more ontologies in their specification, and the meta-information used in specifying an MREF is mapped to views involving keyword-based, attribute-based, and content-based specifications involving various types of metadata of heterogeneous digital media. Specification and processing based on information correlation can be easily integrated with the Web technology. For example, MREF could be used anywhere a hypermedia link (HREF) is used, and its specification and processing can be supported using an RDF and XML-based infrastructure. However, many challenges remain in extending the current proposals to include non-standard resources such as datasets and procedures, integrating information correlation representation and processing with context and context mediation, and processing them efficiently in a very large information space.

Context is commonly conceived to be constructed partly on the basis of mutually accepted propositions (beliefs) (Ouksel 1999). These mutual beliefs are expected to bear on establishing shared ontologies and regulate domain-specific collaboration. While the metaphor of constructing a context appropriately connotes activity, we proposed in (Ouksel 1999) to supplement that with another metaphor connoting an even more dynamic development: interacting agents negotiate contexts. This is essential in an environment of continuously evolving semantics. Clearly, we believe this area will continue to be an important research challenge.

Reasoning about context mappings occurs generally under incomplete information (Ouksel and Naiman 1994). The robustness of semantic interoperability solutions will depend to a large extent in their ability to resolve conflicts in less than ideal situations such as these.

About this special section

Given a possibly broad interpretation of what is semantics, our emphasis has been to focus on real-world semantics rather than semantics of formal representations or systems (e.g., semantics associated with a first order logic or formal axiom system). That is, semantics related to mapping of objects in the model or computational world onto the real world, or the issues that involve human interpretation, or meaning and use of data or information, are of more interest. Items of specific interest include:

- use of domain specific metadata, domain specific ontologies and context to achieve semantic interoperability
- semantics of visual, scientific and engineering data
- fundamental issues in representation and reasoning about real world semantics to achieve semantic reconciliation, identify relationships or measure semantic proximity
- semantic reconciliation amongst structured, semi-structured and multimedia information sources; semantic reconciliation to resolve spatial and temporal conflicts
- theories for supporting dynamic integration of autonomous and heterogeneous information sources with possibly evolving and incompatible internal semantics; semantic negotiation and reconciliation tools in environments characterized by incomplete and uncertain information
- semantic protocols to support intelligent and query-directed integration of information where semantics are viewed as a matter of continuous negotiation and evolution; coordination and search mechanisms to support semantic reconciliation
- semantic interoperability challenges in specific domain (such as those mentioned above or the collaborative domains such as digital earth, etc.)

The call for papers for this special review received excellent response. From among 35 submissions of mostly short descriptions of proposed papers, we selected 9. In this selection process, we preferred the following key criteria of relevance:

- Clearly deal with semantics-- define their definition of semantics, its use in supporting interoperability, integration or cooperation. Furthermore we are interested in "real world"

semantics involving human interpretation and use of information, and the role or use of ontologies, contexts, and other tools that help in capturing and reasoning about semantics. Consequently, if we were to look at a three layer architecture where the layers deal with the issues of data (including syntax/structure/representation, and the corresponding techniques such as generating/using wrappers), metadata, and semantics, then we are less interested in the first two layers.

- Clearly involve global scale, as is possible with the Internet-based infrastructure.

- Involve a broader variety of (heterogeneous) media and information, as well as independently managed (autonomous) components and information sources. As a corollary, we are less interested in approaches and architectures that are variants of federated and multidatabase systems or mediator architectures whose components are primarily structured databases. As a corollary, we have a preference for issues involved in information brokering over a broad variety of distributed, heterogeneous and autonomous information (res)sources.

Overview of the special section

The current section includes a variety of articles on semantic interoperability and represents an interesting mix of applications and conceptual approaches. The first article "Semantic Integration of Environmental Models for Application to Global Information Systems and Decision-Making" by Scott Mackay, discusses the issue of weak and poorly defined semantics in spatially distributed environmental models. He concludes that many issues associated with weak model semantics can be resolved with the addition of self-evaluating logic and context-based tools that discover and exhibit semantic weaknesses to the end-user.

The second article "Semantic and Pedagogic Interoperability Mechanisms in the ARIADNE Educational Repository" by E. Forte et al., reports on the principles underlying the semantic and pedagogic interoperability mechanisms in an educational and training application. This is an example where semantic and pedagogic principles underlying the construction of the repository are mainly empirical and stem from pragmatic considerations.

The third article "Unpacking The Semantics of Source and Usage To Perform Semantic Reconciliation in Large Scale Information Systems", by Ken Smith and Leo Obrst, discusses some the semantic interoperability challenges in the United States Department of Defense (DoD) and shows that despite the innovation in integration infrastructures these challenges persist. An architecture to support inference of the semantic context of attributes is presented.

The fourth paper "Semantic Video Indexing: Approach and Issues" by Arun Hampapur discusses effective indexing and retrieval in video indexing systems. It examines the issues involved in the design of domain specific video management systems and concludes by emphasizing the importance of semantic knowledge models to insure more sophisticated patterns of querying and browsing video. While this application is relatively new, it raises important semantic interoperability questions.

The fifth paper "Contextualizing the Information Space in Federated Digital Libraries" by M. P. Papazoglou and J. Hoppenbrouwers presents an approach to semantically partition the information space and proposes facilities to contextualize the information available in subject-specific categories.

The sixth paper " Dynamic Service Matchmaking Among Agents in Open Information Environments" by Katia Sycara, Matthias Klusch and Seth Widoff proposes a common language for interacting heterogeneous software agents to describe their capabilities and requests. This common language allows agents in a distributed heterogeneous environments to specify local application domain knowledge and requests and other local information. In turn, this knowledge is used to resolve both syntactic and semantic conflicts which arise during the matchmaking process, and construct filters.

The seventh paper " Semantic Integration of Semistructured and Structured Information Sources" by Sonia Bergamaschi et al. describes the MOMIS (Mediator envirOnment for Multiple Information Sources) approach to the integration and query of multiple, heterogeneous information sources, containing semistructured and structured data. It

focuses on ways of capturing and reasoning about semantic aspects of metadata descriptions. It relies on a description logic kernel language to support analysis of source descriptions and the generation of a consistent common thesaurus, which is in turn for semantic reconciliation.

The eighth paper "Agent-Based Semantic Interoperability in InfoSleuth" by Jerry Fowler et al. Describes EDEN (Environment Data Exchange Network) which applies InfoSleuth -- a distributed agent architecture that addresses the semantic interoperability among information sources and analytical tools within diverse application domains via the use of ontologies -- to environmental information resources provided by agencies located in several states.

Finally, the ninth paper "Semantic Interoperability in Information Services: Experiencing with CoopWare" by Avigdor Gal proposes a coordination mechanism to serve as the basis for a generic architectures for information services. This architecture generates a domain model of the application using a reactive approach. The main idea is to utilize this mechanism to dynamically support the updating of ontologies as the semantics of the data sources change.

Acknowledgements

We would like to thank Tarcisio Lima for his assistance in managing the process of preparing this special section and for providing reviews and comments of the submissions.

Arens Y, Knoblock C A, Shen W 1996 Query reformulation for dynamic information integration. In Wiederhold G (ed) *Intelligent Integration of Information.* Kluwer Academic Publishers: 11-42

Beech D 1997 Data semantics on the information superhighway. In Meersman R, Mark L (eds) *Database Application Semantics.* Chapman and Hall.

Daniel R, Lagoze C, Payette S 1998 A metadata architecture for digital libraries. Proceedings of the Forum on Research and Technology Advances in Digital Libraries (ADL'98), Santa Barbara: 276-288

Dao S, Perry B 1996 Information mediation in cyberspace: scalable methods for declarative information networks. In Wiederhold G (ed) *Intelligent Integration of Information.* Kluwer Academic Publishers: 43-62

Genesereth M, King R (eds) 1995 *Reference Architecture,IntelligentIntegration of Information.* Stanford University and University of Colorado. *http://logic.stanford.edu/architecture/reference.html*

Gruber T 1991 The role of a common ontology in achieving sharable, reusable knowledge bases. *Proceedings of the 2nd International Conference on Principles of Knowledge Representation and Reasoning, Cambridge:* 601-602

Guarino N 1998 Formal ontology and information systems. *Proceedings of the 1st International Conference on Formal Ontology in Information Systems [FOIS'98], Torino*: 3-15

Huhns M, Singh M 1997 Ontologies for agents. *Internet Computing* 1(6): 81-83

Kashyap V, Sheth A 1996 Schematic and semantic similarities between database objects: a context-based approach. *The Very Large Databases Journal* 5(4): 276-304

Kashyap V, Sheth A 1998 Semantic heterogeneity in global information systems: the role of metadata, context and ontologies. In Papazoglou M, Schlageter G (eds) *Cooperative Information Systems: Current Trends and Directions.* Academic Press: 139-178

Khang J, McLeod D 1998 Dynamic classificational ontologies: mediation of information sharing in cooperative federated database systems. In Papazoglou M, Schlageter G (eds) *Cooperative Information Systems: Current Trends and Directions.* Academic Press: 179-203

Lee J, Madnick S, Siegel M 1996 Conceptualizing semantic interoperability: a perspective from the knowledge level. *International Journal of Cooperative Information Systems* 5(4)

Manola F 1998 *Towards a Web Object Model.* Object Services and Consulting, Inc. http://www.objs.com/OSA/wom.htm

Meersman R 1997 An essay on the role and evolution of data(base) semantics. In Meersman R, Mark L (eds) *Database Application Semantics.* Chapman and Hall

Mena E, Illarramendi A, Kashyap V, and Sheth A 1999 OBSERVER: An Approach for Query Processing in Global Information Systems based on Interoperation across Pre-existing Ontologies, Distributed and Parallel Databases Journal

Naiman C. and Ouksel, A. M. 1995 A classification of Semantic Conflicts in Heterogeneous Information

Systems *Journal of Organizational Computing* 5(2), 167-193, 1995.

Paepcke A, Chang C, Garcia-Molina H, Winograd T 1998 Interoperability for digital libraries worldwide. *Communications of the ACM* 41(4): 33-43

Ouksel, A. M., Naiman, 1994 Coordinating Context Building in Heterogeneous Information Systems *Journal of Intelligent Information Systems* 3,1,151-183.

Ouksel, A. 1992 Semantic Mechanisms for Cooperation in Heterogeneous Database Systems. *International IEEE Conference on Man and Cybernetics*, October .

Ouksel, A., Iqbal A. 1999 Ontologies are not the Panacea in Data Integration: A Flexible Coordinator to Mediate Context Construction. *Distributed and Parallel Databases,* 7, 1-29.

Ouksel, A. 1999. A Framework for a Scalable Agent Architecture for Cooperating Heterogeneous Knowledge Sources. *Intelligent Information Agents: Cooperative, Rational and Adaptive Information Gathering in the Internet.* M. Klusch (Ed.), Chapter 5, Springer.

Shah K, Sheth A 1998 Logical information modeling of Web-accessible heterogeneous digital assets. Proceedings of the Forum on Research and Technology Advances in Digital Libraries (ADL'98), Santa Barbara: 266-275

Sheth A 1998, Changing Focus on Interoperability in Information Systems: From System, Syntax, Structure to Semantics, in Interoperating Geographic Information Systems, M. F. Goodchild, M. J. Egenhofer, R. Fegeas, and C. A. Kottman (eds) Kluwer Publishers

Sheth A, Kashyap V 1993 So far (schematically) yet so near (semantically). In Hsiao D, Neuhold E, Sacks-Davis R (eds) 1993 *Interoperable Database Systems (IFIP Transaction A-25, Proceedings of DS-5)* North-Holland