Semantic Web Workshop: Models, Architectures and Management
September 21, 2000, Lisbon, Portugal
(held in conjunction with ECDL 2000)

Panos Constantopoulos, Vassilis Christophides, Dimitris Plexousakis,
University of Crete, Heraklion, Crete, Greece and
Institute of Computer Science, Foundation for Research and Technology – Hellas
{panos, christop, dp } @ics.forth.gr

Motivation

In 1998, the World-Wide Web Consortium (W3C) inaugurated a research initiative centered on the idea of providing semantics for and facilitating the extraction of knowledge from the World-Wide Web. The Semantic Web is a vision of the creator of the WWW, Tim Berners-Lee, who describes it as "a web of data, documents, or portions of documents, that can be processed directly or indirectly by machines" not just for display purposes, but for automation, integration and reuse across various applications. The primary goal of the Semantic web is to define infrastructure, standards and policies facilitating an explicit description of the *meaning* of Web resources that can be processed both by automated tools and people. This effort towards the next evolution step of the Web, has given rise to a large number of research problems that relate to models, architectures, applications and services for the Semantic Web. The following is a list of research issues that were put forth as themes for soliciting workshop submissions.

- Formal Foundations of Web Metadata Standards
- Semantic Interoperability Frameworks
- Information and Services Brokering Architectures
- Metadata Creation, Extraction and Storage
- Query Languages for the Semantic Web
- Distributed Inference Services
- Digital Signatures and Web of Trust
- Advanced Resource Discovery Interfaces
- Automated Classification of Web Resources
- Superimposed Web Resource Annotation & Recommendation Tools
- Personalization and Intellectual Property Rights
- Semantic Web Applications: Knowledge Portals, Electronic Commerce

The objective of the workshop was the creation of a forum for presenting research results in developing infrastructure for the Semantic Web and for enabling and fostering interaction amongst international researchers. The collocation of the workshop with the European Conference on Digital Libraries broadened the intended scope of the workshop and attracted participation and interaction from industry in addition to the academic and research communities. The workshop's audience comprised researchers and practitioners in the areas of databases, intelligent information integration, knowledge representation, knowledge management, information retrieval, metadata, web standards, digital libraries and others.

The workshop was organized as a post-conference one-day workshop at ECDL 2000 in Lisbon, Portugal. Panos Constantopoulos chaired the workshop committee with Vassilis Christophides and Dimitris Plexousakis as Program Committee Co-Chairs. A total of 29 papers were submitted to the workshop. Each paper was peer-reviewed by at least two referees. Despite the overall high quality of the submissions, only 9 papers were accepted by the program committee for presentation at the single-day event. Overall, the workshop drew considerable attention at ECDL 2000: 63 registered participants from 22 countries. The workshop was sponsored by ERCIM (the European Research Consortium on Informatics and Mathematics) and the DELOS Network of Excellence on Digital Libraries.

Keynote Address

The keynote address entitled "*Semantic Web and Information Brokering: Opportunities, Early Commercializations, and Challenges*" was delivered by *Amit Sheth* (University of Georgia and Taalee Corp). Sheth characterized semantics as the next step in the evolution of the WWW and stressed the importance of semantically organized information for supporting ubiquitous, powerful, accurate and efficient access to this information. Sheth also reviewed proposals for semantic interoperability frameworks such as the DAML (DARPA Agent Mark-Up Language), the Oingo family of tools for defining concepts and extracting knowledge from large databases, as well as several scenarios on learning on the Web. He moved on to present the semantic services provided by Taalee, including semantic categorization, cataloguing, search, personalization and targeting. Although given mainly from a commercial perspective, Sheth's presentation made a clear statement about the importance of semantic enrichment in enabling information brokering on interoperable multi-database systems. Terminology and language transparency, comprehensive metadata management, context-sensitive information processing and semantic correlation were characterized as the basis for enabling the symbiosis of semantic information brokering and the Semantic Web.

Technical Paper Sessions

The technical paper presentations were organized in three sessions addressing Semantic Interoperability Frameworks, Web Metadata Models & Standards, and Applications & Systems for Community Webs. Three papers were presented in each of the sessions.

Session I: Semantic Interoperability Frameworks

The first paper of this session, authored by *Lois Delcambre* and *Shawn Bowers*, was entitled "*Representing and Transforming Model-Based Information*." The paper advocates the use of a mapping formalism for converting superimposed information from one representation scheme to another. Superimposed information consists of data placed over existing information sources for annotating, supplementing and interconnecting underlying information. In this manner, tools developed for each representation (XML, RDF, Topic Maps) can be used by simply converting the information from its original form to that required by the tool. The paper presents a generic representation scheme for model-based information, using a metamodel expressed in RDF and RDF Schema, as well as a set of mapping rules for information representation. The metamodel proposed comprises a minimal set of abstractions for the model-based superimposed information which capture structural definitions within models, their relationships and associated constraints. The paper proceeds by defining a visual representation of models defined by the metamodel using a subset of UML. A technique for manual specification of mapping rules as production rules is proposed. The rules are defined over the triple representation of RDF in a logic-based language. Mappings are performed by invoking mapping functions for conversion, and model and schema extraction. In this manner, inter-model, inter-schema and model-to-schema mappings (and mixtures of them) can be implemented.

The second paper of this session, authored by *Sergey Melnik* and *Stefan Decker*, was entitled "*A Layered Approach to Information Modeling and Interoperability on the Web*." The focus of the paper is interoperability between different information models. Motivated by ideas in applications internetworking, a layered approach towards achieving information interoperability is suggested. The main idea is the employment of a - so called - "object layer" as a bridge between syntactic and semantic variations of existing incompatible information models on the Web. The layered view of Web-enabled information models consists of the syntax, object and semantic layers, each of which may in turn consist of several sub-layers corresponding to specific modeling features of the information model. The object layer, which is the main focus of the paper, provides as a basic functionality the manipulation of objects and their interrelationships. It consists of five sub-layers dealing with object identity, relationships, basic typing, reification and ordering. The authors provide a review of the logical implementation alternatives of the sub-layers in different data modeling languages (OEM, RDF/S, SHOE, UML, OIL) and comment on the design of APIs for navigational and declarative access to information, and mappings to the syntactic layer.

The last paper in this session, authored by *Jutta Eusterbrock*, was entitled "*Knowledge Mediation in the WWW Based on Labelled DAGs with Attached Constraints*." In this paper, Directed Acyclic Graphs (DAGs) are proposed as a natural formal model for capturing syntactic and semantic properties of Web data, ontologies, their constraints and operations. The problem of mediating heterogeneous information sources is viewed as a synthesis problem that requires taking into account constraints on the structure and content of Web resources as well as interaction constraints. The proposed mediation architecture is based on the SEAMLESS framework for knowledge integration and transformation. It employs an intermediate context-dependent mediator knowledge base that incorporates a shared ontology, XML DTDs and viewpoints. The latter are used for the purpose of creating mediator knowledge bases that combine a shared ontology and DTDs. Lifting rules are employed in order to create an abstraction of data from local repositories and enable their integration with others from different sources. Thus, users or software agents may query the mediator knowledge base without having to know the exact location, the format or the specific context of the underlying data. Precise requests are generated using knowledge-based reasoning and constraint solving using a logical representation of graph terms.

Session II: Web Metadata Models and Standards

The first paper of this session, authored by *Steffen Staab et al.* was entitled "*An Extensible Approach for Modeling Ontologies in RDF(S)*." The paper's theme is the modeling of ontologies in RDF(S) by incorporating axioms at the knowledge level. The described approach is open and extensible, in that it permits users or user-communities to specify axioms pertaining to their domain of interest, whilst preserving and reusing the core RDF(S) semantics. Domain and application specific axioms are produced by means of translating RDF(S) axiom specifications into various target systems that employ them in inferencing. Axioms are categorized to allow for a more concise definition of their semantics and to serve the purpose of interchangeability and adaptability. The initial axiom specification takes place in RDF(S) independently of the target systems. Particular axioms pertaining to specific inference engines are defined as instances of the axiom schemata specified in RDF(S). As an example of the applicability of this methodology, the authors present the translation of a small set of axioms in F-logic.

The second paper of this session, authored by *Jeen Broekstra et al.* was entitled "*Adding Formal Semantics to the Web: Building on Top of RDF Schema*." The paper builds on the premise that an additional logical layer is needed on top of RDF Schema, in order to provide semantics for and enrich RDF Schema. The authors describe the ontology language OIL as an extension to RDF Schema and provide an RDF schema definition of the language primitives. OIL provides most of the modeling primitives found in frame-based languages and provides reasoning support for subsumption and class consistency checking. The paper demonstrates how OIL class definitions can be written in RDF by reusing existing RDFS constructs and extending RDF(S) by constructs such as *class expressions* (i.e., extensions to simple class names), Boolean operators and slot definitions and constraints. The resulting extension permits RDF(S)-aware systems to process OIL ontologies and OIL-aware systems to take advantage of the formal semantics and reasoning support.

The last paper in this session, authored by *In-Young Ko et al.*, was entitled "*Semantically-Based Active Document Collection Templates for Web Information Management Systems*." The authors describe techniques for representing semantics of both document collections and information management services operating on them. These techniques allow users to create semantic descriptions for the collections and services and render them interoperable with other services that (re)use the same descriptions. Document collection semantics are divided into two independent types: content and organization structure. Services, including analysis and visualization, are supported by *active document templates* that maintain semantic relations between document collections and which can dynamically be modified and instantiated to generate collections for similar tasks. The proposed framework enables reusability and exchangeability of templates, thus serving to improve performance of large-scale information management tasks.

Session III: Applications and Systems for Community Webs

The first paper of this session, authored by *Jose Maria Abasolo and Mario Gomez* was entitled "*MELISA: An Ontology-Based Agent for*

Information Retrieval in Medicine." The paper describes a prototype of an ontology-based information retrieval agent assisting a user in formulating queries using different ontologies and query models, and in aggregating and filtering query results generated by different information resources. Three levels of abstraction are proposed: (1) a high-level consultation query, employing generic terms and categories, (2) a conceptual query, generated by connecting the consultation query with terms in the ontologies, and (3) a specific query, i.e., a set of low-level, database-dependent queries. A decomposition process translates a consultation query into a number of conceptual queries, one for each of the categories selected. The conceptual queries are in turn decomposed into a set of specific queries using the keywords and filters specified in the higher levels. A filter and combination mechanism joins and filters the returned answers according to relevance ratings and scores assigned based on the query characteristics.

The second paper of this session, authored by *Stanislaw Ambroszkiewicz* was entitled "*Semantic Interoperability in Agentspace: Proposal for Agent Interface to Environment* ." The paper focuses on extending basic semantic interoperability infrastructure for interaction and knowledge exchange among agents. The author proposes a formal specification of the agents' interaction structure using a mobile agent platform called Pegaz. The proposed interface serves as a means to achieve semantic interoperability. The main idea put forth in the paper is that a generic representation of world structure, common to all interacting agents, serves as the way to exchange meaning by having a common perception of the environment. A 3-layered interface is proposed, comprising a functionality layer, standardizing the basic interaction types, a representation layer, specifying formally the agents' perception structure as determined by the functional layer, and a language layer, establishing the meaning interchange syntax. The representation layer establishes an ontology core that formalizes the structure of the world, realized by the functionality layer.

The last paper in this session, authored by *John Pierre* was entitled "*Practical Issues for Automated Categorization of Web Sites*." This last technical presentation focused on the problem of automated text classification for web sites by a so-called "targeted spidering" approach, involving metadata extraction and opportunistic crawling of semantic hyperlinks. The importance of automated text classification is warranted by the vast amount of constantly changing web resources. The paper consists of a quantitative analysis of several thousands of randomly chosen domains, based on the meta-tags employed in the documents. The analysis results in a number of "good" text features that the authors argue should be employed by text classification methods. The opportunistic spidering method attempts to extract useful text from metatags and titles of web pages and then follow existing hyperlinks. The link-following strategy is based on an ad-hoc frequency analysis. Such an approach could benefit from a formalized description of semantic relationships among web pages. The paper also presents results from experiments, conducted in an industrial domain. The results show that metatags are the best source of quality text features as compared to the body of text contained in a web page. The paper concludes with a proposal for building a targeted automated web categorization system relying on knowledge gathering, targeted spidering, training and classification.

Invited Talk

Hans Georg Stork of the European Commission gave a presentation on the IST Work Programme Vision and Priorities on the Semantic Web. His talk was a precursor to the official launch of the Semantic Web Action line under the World-Wide Web Consortium. Semantic Web technologies constitute the focus of IST Programme – Key Action III for year 2001 under the theme of Information Access, Filtering, Analysis and Handling. Dr. Stork identified three main axes of research towards the Semantic Web, namely *formalizing*, *grounding*, and *acting*, referring to: (a) methods and tools for coding/structuring digital content and for defining and declaring its semantics, (b) methods and tools for the derivation of semantic attributes of web-based content, and (c) semantics-based tools for knowledge/resource discovery, intelligent filtering and profiling, transactions, and querying.

Summary

The goal of this first workshop was to promote the formation of a multidisciplinary community working on the theory and implementation of the Semantic Web. The keynote address and the technical papers presented in the Workshop focused on some of the issues emerging as research themes around the idea of the Semantic Web. Figure 1 presents the workshop attendees' view of the identified core research topics, only few of which (see shadowed boxes) has been addressed at the single-day event.

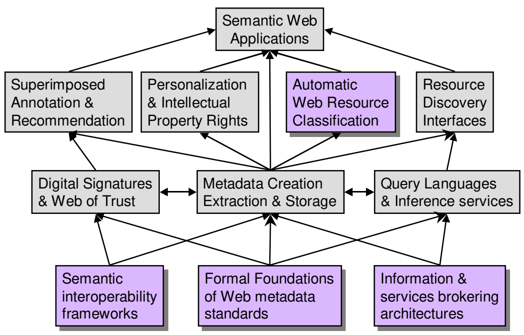

Figure 1: An emerging agenda for the Semantic Web

A consensus reached at the end of the workshop was that there is sufficient interest in the international community of researchers and practitioners for a continuation of the thematic workshop on the theory and implementation of the Semantic Web. In fact, the Second International Workshop on the Semantic Web was recently held at WWW10 in Hong-Kong.

Program Committee

General chair

Panos Constantopoulos (ICS-FORTH & Univ. of Crete, Greece)

Program chairs

Vassilis Christophides (ICS-FORTH)

Dimitris Plexousakis (ICS-FORTH & Univ. of Crete, Greece)

Program Committee

Bernd Amann (CNAM-Paris, France)

Thomas Baker (GMD, Germany)

Manolis Koubarakis (Technical University of Crete, Greece)

Alain Michard (INRIA, France)

Marie-Christine Rousset (Univ. of ORSAY, France)

Michel Scholl (CNAM-Paris, France)

Amit Sheth (Univ. of Georgia & Taalee Corp. USA)

Anne-Marie Vercoustre (INRIA, France)

Acknowledgements

The workshop's organizing committee would like to thank the organizers of ECLD-2000 for their invaluable assistance in holding this successful event.