Dagstuhl-Seminar: Semantics for the WWW

Dieter Fensel[1], Jim Hendler[2], Henry Lieberman[3], and Wolfgang Wahlster[4]

[1] Vrije Universiteit Amsterdam, De Boelelaan 1081a, 1081 HV Amsterdam, The Netherlands,
phone/fax: +31-(0)20-872 27 22, dieter@cs.vu.nl, http://www.cs.vu.nl/~dieter.

[2] Department of Computer Science, University of Maryland, College Park, MD 20742, USA,
phone: (301) 405-2662, fax: (301) 405-6707, hendler@cs.umd.edu, http://www.cs.umd.edu/~hendler.

[3] MIT Media Laboratory, 20 Ames St. 305 A, Cambridge, MA 02139 USA,
phone: (+1-617) 253-0315, fax: (+1-617) 253-6215, lieber@media.mit.edu, http://lieber.www.media.mit.edu/people/lieber

[4] DFKI GmbH, Stuhlsatzenhausweg 3, D-66123 Saarbrücken, Germany
phone: +49 681 302 5252, fax: +49 681 302 5341, wahlster@dfki.de, http://www.dfki.de/~wahlster/

Abstract. Currently computers are changing from single isolated devices to entry points in a world wide network of information exchange and business transactions called the World Wide Web (WWW). Therefore support in data, information, and knowledge exchange becomes the key issue in current computer technology. The WWW has drastically changed the availability of electronically available information. However, this success and exponential grow makes it increasingly difficult to find, to access, to present, and to maintain the information of use to a wide variety of users. In reaction to this bottleneck many new research initiatives and commercial enterprises have been set up to enrich available information with machine processable semantics. Such support is essential for "bringing the web to its full potential". This semantic web will provide intelligent access to heterogeneous and distributed information enabling software products (agents) to mediate between the user needs and the available information sources. The report summarizes a Dagstuhl seminar on these topics that was held during March 2000 in Dagstuhl, Germany.

1 Introduction

The World-wide Web (WWW) has drastically changed the availability of electronically available information. Currently there are around 300 million static documents in the WWW which are used by more than 100 million users internationally. In addition, this number is growing astronomically. In 1990, the WWW began with a small number of documents as an in-house solution for around thousand users at CERN. By 2002, the standardization committee for the WWW (called W3C) expects around a billion web users and a even higher number of available documents. However, this success and exponential grow makes it increasingly difficult to find, to access, to present, and to maintain the information of use to a wide variety of users. Currently, pages on web must use representation means rooted in format languages such as HTML or SGML and make use of protocols that allow browser to present information to human readers. The information content, however, is mainly presented by natural language. Thus, there is a wide gap between the information available for tools that try to address the problems above and the information kept in human readable form.

- **Searching for information:** Already, finding the right piece of information is often a nightmare. One gets lost in huge amounts of irrelevant information and may often miss the relevant ones. Searches are imprecise, often returning pointers to many thousands of pages (and this situation worsens as the web grows). In addition, a user must read retrieved documents in order to extract the desired information -- so even once the page is found, the search may be difficult or the information obscured. Thus, the same piece of knowledge must often be presented in different contexts and adapted to different users' needs and queries. However, the web lacks automated translation tools to allow this information to automatically be transformed between different representation formats and contexts.
- **Presenting information:** A related problem is that the maintenance of web sources has become very difficult. Keeping redundant information consistent and keeping information correct is hardly supported by current web tools, and thus the burden on a user to maintain the consistency is often overwhelming. This leads to a plethora of sites with inconsistent and/or contradictory information.
- **Electronic commerce:** Automatization of electronic commerce is seriously hampered by the way information is currently presented. Shopping agents use wrappers and heuristics to extract product informations from weakly structured textual information. However, development and maintenance costs are high and provided services are limited.

There is an emerging awareness that providing solutions to these problems requires that there be a machine understandable semantics for some or all of the information presented in the WWW. Achieving such a semantics requires:

- Developing languages for expressing machine understandable meta information for documents.
- Developing terminologies (i.e., name spaces or ontologies) using these languages and making them available on the web.
- Integrating and translating different terminologies.
- Developing tools that use such languages and terminologies to provide support in finding, accessing, presenting and maintaining information sources.

Developing such languages, ontologies, and tools is a wide ranging problem that touches on the research areas of a broad variety of research communities. Therefore this seminar brought together colleagues from these different research communities (who would typically not meet at area conferences or other workshops that are more methodologically driven). These include researchers in the areas of databases, intelligent information integration, knowledge representation, knowledge engineering, information agents, knowledge management, information retrieval, meta data, web standards (RDF, XML, XML-QL, XSL), and others. The goal of this meeting in Dagstuhl, March 19-24, 2000, was to bring together the scientists and technologists working in these areas, and to thus allow the exchange of information about emerging tools and techniques.

The contents of the seminar was organized as follows. First, we discussed a number of arising new web standards that should improve the representation of machine processable semantics of information. Second, we discussed the use of ontologies for representation of semantics (in the sense of formal and real-world semantics). Third, these semantic annotations allow automatization in information access and task achievement. Therefore, we discussed intelligent information access based on these semantic annotations. Forth, we discussed a number of applications of these new techniques and had a number of exiting demonstrations. Last but not least we had some workshops dealing with emerging issues during the seminar.

The presentations are available at http://www.semanticweb.org/events/dagstuhl2000/.

2 New web standards

- Ora Lassila: RDF: A Frame System for the Web

  RDF is a new standard for web metadata [Lassila 1998], published by the World Wide Web Consortium (W3C). It is intended as a primitive knowledge representation system for internet-based applications, and as such serves as the basis for W3C's vision of the "Semantic Web". RDF consists of a simple data model, akin to semantic networks, layered on top of an XML-based syntax. It also introduces an object-oriented, extensible type system and some meta-constructs (namely container types and higher-order statements).
  We have encountered some difficulties in explaining RDF to the general web developer community. RDF's syntax may be somewhat cumbersome, but the bigger and deeper issue is the general difficulty of modeling. It is our belief that if RDF was presented as a frame-based representation system, it would be easier to understand. Frame systems, as structural modeling tools, are generally well accepted and are easy to comprehend (some terminology translations are necessary, though: "frame" vs. "resource", "slot" vs. "property", etc.). On top of this we can then layer some logic capabilities; at least, this should take the form of some type of description logic.

- Henry S. Thompson: Internet-based Application Architectures for the 21 Century: The Role of XML

  There is evidently a significant convergence between two important technologies: markup for the World Wide Web and databases. The Internet is just beginning to make an impact on the database world, and the fact that data as much as if not more so than documents will use XML to travel the Internet is just beginning to significantly influence the design and development of the XML family of standards.

  XML has defined a transfer syntax for tree-structured documents; Many data-oriented applications are being defined which build their own data structures on top of an XML document layer, effectively using XML documents as a transfer mechanism for structured data

- Stefan Decker: Putting things together

  The WWW is a very heterogeneous information collection, ranging from Chemical Process Descriptions[1], over Probabilities and Fuzzy Theory[2], to classifications schemes for minerals[3]. Currently this information is represented in text-Form, deploying HTML for presentation. In the next-generation Web, the Semantic Web, all this information is available for automated agents, doing tasks for their human users. However, there is nothing like a "one-size-fits-all" Knowledge Representation Language: requirements for Knowledge Representation Languages are as diverse as the Knowledge that needs to be represented. Hence, focusing on

1. http://www.gulfchem.com/Process.htm
2. http://www.dbai.tuwien.ac.at/marchives/fuzzy-mail98/1245.html
3. http://www.usoe.k12.ut.us/curr/science/core/7thgrd/integrated/classifcation/sciber/rock.htm

creating "the" Web-Representation Language is not suitable: instead one has to cope with various, heterogeneous Knowledge Representation mechanisms and services on the Web.

To be able to take advantage about from the information available in the Web, establishing interoperability between various Services in a cost effective way is the most important goal. We propose a framework for establishing interoperability on the Web deploying RDF as the intermediate data-model between services: services are exchanging information using RDF in their own vocabulary (the Knowledge Representation Language). Interoperability is again established using RDF, defining mapping rules for vocabularies. Basic services can be composed to more sophisticated ones be declarative specifications, available on the web for everyone to deploy. We believe that this mechanism will be the integral part of the Semantic Web.

- Massahiro Hori: Annotation of Web-Content for Transcoding

  Users are increasingly accessing the Internet from Web-enabled personal devices. Since such devices do not have the same rendering capabilities as desktop computers, it is necessary for Web content to be adapted, or transcoded, for proper presentation on a variety of client devices. In this talk, I briefly introduce a framework of external annotation, in which existing HTML documents are associated with content adaptation hint as separate annotation files written with XML/RDF. An annotation-based transcoding system is then presented with particular focus on the authoring-time integration between a WYSIWYG authoring tool and a transcoding module. Finally, a short demonstration is given an example of content adaptation using a page-splitting module for small-screen devices.

3 Ontologies

- Guus Schreiber: Requirements for Ontology Specification Languages

  Ontologies are increasingly seen as an important vehicle for describing the semantic content of web-based information sources. The notion of ontology has been the subject of many debates over the past years. In this talk we provide a strong and a weak definition of "ontology". The strong version reads: "an ontology is an explicit specification of a shared conceptualization that holds in a particular context". The weaker notion of ontology leaves out the word "explicit" and thus also includes corpora such as large thesauri (e.g. Art and Architecture Thesaurus) in which the specification has to be entangled form the corpus (e.g, hierarchy structure). Several types of ontologies exist: domain-oriented ontologies (specific for a particular device), task-oriented ontologies (specific for a certain problem context, such as instruction) and generic ontologies (top-level categories, in the spirit of Aristotle). Typically, we need several a mix of ontology types for particular applications. We illustrate this with an example in which a technical manual is decomposed and indexed for instructional purposes.

  A number of ontology specification languages exist, such as KIF, Ontolingua, and LOOM. More general modelling languages are also used, e.g. Express and UML. All share a number of features, in particular classes, generalization, and relations. Often however, additional expressivity is needed. We mention: multiple subclasses, relation/attribute distinction,

aggregation, relations as classes, a constraint language, class/subclass semantics, modularization mechanisms, predefined data types, ontology-mapping mechanisms, and sloppy class/instance distinctions. We present a priority list for inclusion of such features in a specification language for semantics of web information. We also discuss how well the current version of the web standards RDF and RDFS can express ontological constructs. We conclude with a wishlist for managing ontologies in RDF/RDFS. This list includes a graphical representation (e.g., UML-based), editing/manipulation tools (e.g., Protege-2000), convincing examples and applications, as well as methods and guidelines for RDF/RDFS usage in ontology specification.

- Frank van Harmelen & Ian Horrocks: OIL: A Proposal for an Ontology Interchange Language

  Currently computers are changing from single isolated devices to entry points in a world wide network of information exchange and business transactions. Therefore, support in data, information, and knowledge exchange becomes the key issue in current computer technology. Ontologies provide a shared and common understanding of a domain that can be communicated across people and application systems. Ontologies will play a major role in supporting information exchange processes in various areas. A prerequisite for such a role is the development of a joint standard for specifying and exchanging ontologies. The purpose of this talk is precisely concerned with this necessity. We present the Ontology Interchange Language OIL which is a proposal for such a standard. It is based on existing proposals such as OKBC, XOL and RDF and enrich them with necessary features for expressing rich ontologies. The talk presents motivation, underlying rationale, modeling primitives, syntax, semantics, and tool environment of OIL. With OIL, we want to make a proposal opening the discussing process that may lead to a useful and well defined consensus of a large community making use of such an approach.

- Jeff Hefflin: SHOE: A Knowledge Representation Language for the Web

  The Internet is an information resource with virtually unlimited potential. However, this potential is relatively untapped because it is difficult for machines to perform useful processing on this information. The task of locating relevant pages is time consuming, and more complex tasks such as comparing resources at different web sites are virtually impossible. Some have championed XML as a panacea for these problems, but XML is only a partial solution. True machine understandable knowledge is needed to use the Internet intelligently. However traditional knowledge representation cannot solve the problem alone, this work must be expanded to deal with the fundamental characteristics of the Web: it represents the potentially inconsistent views of many, it is constantly changing, and it is enormous.

  In the first part of this talk, I present the Simple HTML Ontology Extensions (SHOE) language, a knowledge representation language designed with the needs of the Web in mind. There are two types of SHOE pages: ontologies and instances. Ontologies provide the vocabulary and rules for reasoning about data, while instances contain data and commit to specific ontologies. Ontologies can be extended by new ontologies that add definitions, rules, or alternate terminology, thus enabling interoperability without forcing everyone to commit to a single representation of the world. Ontologies can also be revised in a manner that preserves dependencies of other objects, allowing change to occur as needed.

  In the second part of the talk, I discuss the implementation of the SHOE language, and present an architecture that enables its use. This architecture includes a tool that helps the user

structure knowledge on a web page, a web-crawler that gathers the SHOE information and stores in a knowledge base, and various query tools that can be accessed over the Web. I conclude with a description of the application of this system to two diverse problem domains.

- Christopher A. Welty: Semantics for the web

  The popularity and press surrounding the release of XML has created widespread interest in standards within particular communities that focus on representing content. The dream is that these standards will enable consumers and B2B systems to more accurately search information on the Web within these communities. We believe the expansiveness and diversity of the Web creates a need for a small set of standard semantic primitives that have the same meaning and interpretation across communities. Such a standard set of primitives should take into account existing efforts in ontology, and in e-commerce content standards. We are investigating existing content standards proposals for the Web, and present some basic motivations and very preliminary ideas regarding what such a standard set of semantic primitives could be. I begin with some quotes from the workshop so far, and then present some examples of work in the library and e-commerce domains, and how they might be harmonized.

- Frank Nack: MPEG-7: Semantics for Audio-Visual Media on the Web

  If audio-visual information should be of use as a resource it must allow some degree of interpretation, which can be passed onto, or accessed by a device or computer code. MPEG-7 aims to create a standard for describing these operational requirements. The talk provides an overview on the communicational problems addressed by MPEG-7 based on examples for video, image and audio applications, describes the development of MPEG-7, and discusses the concepts , terminology and requirements.

- Jim Hendler: The DARPA Agent Markup Language (DAML)

  The modern information technology world is a dynamically changing environment with an exponentially increasing ability to create and publish data that rapidly swamps human abilities to process that data into information. Agent-based computing can potentially help us to recognize complex patterns in this widely distributed, heterogeneous, uncertain information environment. Unfortunately, this potential is hampered by the difficulty agents face in understanding and interacting with data that is either unprocessed or in natural languages. The inability of agents to understand the conceptual concepts on a web page, their difficulty in handling the semantics inherent in the outputs of a program, and the complexity of fusing information concept from the outputs of sensors, to name but a few problems, truly keep the "agent revolution" from occurring.

  One potential solution to this problem is for humans to, as it were, meet the computer half way. By using tools to provide mark-up annotations attached to data sources, information can be made available to the agents in new and exciting ways. Going far beyond XML, the goal of this program is to develop a language aimed at representing semantic relations in machine readable ways compatible with current and future Internet technologies. Further, prototype tools will be developed to show the potential of such markups to provide revolutionary capabilities that will change the way humans interact with information. Deployment of such tools to military and intelligence users, and showing the incredible dual use potential of such a technology, caps off the programs goals.

  To realize this solution, Internet markup languages must move beyond the implicit semantic

agreements inherent in XML and community-specific controlled languages, and move towards making semantic entities and markup a primary goal DARPA will lead the way with the development of DARPA Agent markup Language (DAML). DAML will be a semantic language that ties the information on a page to machine-readable semantics (ontology). The language must allow for communities to extend simple ontologies for their own use, allowing the bottom-up design of meaning while allowing sharing of higher level concepts. In addition, the language will provide mechanisms for the explicit representation of services, processes and business models, so as to allow non-explicit information (such as that encapsulated in programs or sensors) to be recognized.

DAML will provide a number of advantages over current markup approaches. It will allow semantic interoperability at the level we currently have syntactic interoperability in XML. Objects in the web can be marked (manually or automatically) to include descriptions of information they encode, descriptions of functions they provide, and/or descriptions of data they can produce. This will allow web pages, databases, programs, models, and sensors all to be linked together by agents that use DAML to recognize the concepts they are looking for. If successful, information fusion from diverse sources will become a reality.

- Deborah L. McGuinness: The “Pull” for Ontologies

  Ontologies have moved beyond the domains of library science, philosophy, and knowledge representation. They are now the concerns of marketing departments, CEOs, and mainstream business. Analyst companies such as Forrester Research report on the critical roles of ontologies in support of browsing and search for e-commerce. One now sees ontologies used as a core controlled vocabulary that is integrated into catalogues, databases, web publications, knowledge management applications, etc. We also see ontologies that have long live spans and end up being distributed in generation and maintenance responsibilities. As the usage of ontologies broadens, the user base broadens and now tool environments become more critical.

  In our work on ontology environments, we have been motivated by the emerging needs for distributed ontology creation and maintenance environments. We consider the task of merging terms in ontologies based on term definitions. We also address the task of analyzing ontologies with an eye towards focusing the user’s attention in areas of ontologies that need modification. The Chimaera ontology environment has been produced from our work. It is a tool that supports ontology merging and analysis and has been used in academic and commercial settings.

- Robert Meersman: Can Ontologies Learn from Database Semantics?

  Databases have become the hugely successful tools they are mostly because they implement so-called "data independence", which for the sake of simplicity one may call the ability to specify and manage data structures *outside* application programs, and consequently to allow management of the (usually large) "populations" of those data structures by specialized, highly efficient software tools. Research and practice in databases have resulted over time in techniques and methodologies for representing information in such data structures that have become quite sophisticated, a fact sometimes overlooked by other research communities. Object-oriented, object-relational and often even "plain" relational database management systems (DBMS) come equipped with a variety of syntactical constructs that permit database- and conceptual schemas to represent objects, subtype taxonomies, and some integrity constraints, derivation rules etc.. A number of methodologies have been developed to assist in the creation of such conceptual database schemas (or “data models” are they are often

imprecisely called), such as EER, ORM, UML, … each supported by a variety of so-called CASE tools. It is important (even essential for our understanding of the relevance of ontologies) to realize that DBMSs and associated CASE environments are geared to providing software solutions a particular *application instance* (e.g. an airline reservation system) rather than for *domains* (e.g. air travel), although of course it will in general be hard to formally define this distinction as domains may be quite small and specialized while application suites may cover a wide range of interrelated objects and functions. In this respect it may be worth to note already that so-called ontologies (that in a way should correspond to domain level knowledge, see further) play a different role in the scheme of things than application-specific data models, an observation not helped by the close resemblance of some ontology specification languages in the literature to latter-day conceptual database schema languages [ISO TR 9003].

Database semantics is the research area covering all aspects of the relationship between the implemented database system and the portion of "reality" it is supposed to render. Intuitively but wholly informally, a "higher" quality of this rendering (i.e. the composition of database schema, the database itself, and the application programs making it available to users) is associated with "more" semantics, i.e. more of the "meaning" of the domain of reality is represented in the database system. Various formalisms exist to make this notion more precise, and this exactly is the place where ontologies will enter the picture. The most common classic formalism, also the most amenable to the use of ontologies, is so-called *declarative* or *Tarski* semantics as may be found in various places in the database and AI literature, as in Reiter's seminal paper [Reiter, 1988] or in the book [Genesereth & Nilsson, 1987]. Essentially it replaces "reality" (the domain) by a *conceptualization*, a mathematical object that typically consists of very elementary constructs such as a set of objects and of (mathematical) relations. Semantics is then formally defined simply as an *interpretation mapping* from the system (or rather from the language describing a system instance in some syntax) to this conceptualization.

The elementariness of the conceptualization constructs is essential, first of all to facilitate agreement about them but also to achieve a semantics which is maximally independent of the chosen database schema language and of the represented domain. It is indeed fundamental to realize that all declarative semantics constitutes a form of *agreement* (since at the very least users, domain experts and designers have to agree on a chosen conceptualization). Since database systems are software solutions for a particular application, such agreement has to be based on a common "perception" of this application's domain. Databases typically do not provide schemas for entire domains, i.e. they do not in general "model reality itself". But if one wants to achieve cooperation, interoperation or just communication between database systems, some form agreement naturally has to be established and formalized about the underlying domain ("reality"). Suitably standardized (and large) ontologies may provide a means for this.

Starting from the almost classical definition of an ontology (as a countable noun) by *T.* Gruber as the specification of a conceptualization, it therefore becomes straightforward to see "pure" ontologies rather as mathematical objects, namely as the domain of the semantic interpretation function under consideration. (Naturally, we shall ultimately have to devise a suitable and convenient computer representation for them, but this is an independent issue.) As argued above, it is therefore important to make the elements of an ontology as simple as possible (even at the price of not modeling a lot of the domain's constraints, rules and other

"defining" properties. We conjecture that most often these properties anyhow will turn out to be application-specific and therefore rather should be represented in an appropriate layer "surrounding" the ontology. To make the distinction explicit, we define an ontology base (or "ontobase") as a large set of lexons, these being 4-tuples of the form $<\gamma\ t_1\ r\ t_2>$ where γ represents a context, t_1 and t_2 are terms and r is a role. The term t_1 is called the headword of the lexon. The precise definitions are left for a forthcoming more complete paper, but the intuition should be fairly obvious. Some details may already be found in [Meersman, 1999]. The pragmatics of an ontology base is that it constitutes a set of "plausible" elementary facts that within a given context (e.g. a set of applications) may hold in the domain under study, implying that no valid application (i.e. database system instance) should be inconsistent with them, i.e. the interpretation of such application should satisfy the ontology base in some well-defined sense. A promising initial formalization of this concept may perhaps be derived from the work of Guarino [Guarino, 1998]. Note that we deliberately exclude derivation rules, constraints and the like from the ontology base, thereby in some cases sacrificing the relative compactness of an intensional representation for a more extensional one but one, we claim, that is easier to agree on.

Evidently the construction (or should we say "growing") of standardizable, hence reusable and dependable computerized ontology bases will not be a mean feat. In the DOGMA project at VUB STARLab[4] we are trying to set up an ontology server in order to assist the gathering and incremental growth of sets of lexons. One important source of lexons coding domain-specific knowledge (as opposed to generic ones occurring in general-purpose lexicons such as Wordnet etc.) will be formed by relational database schemas, yielding an activity best described as ontology mining. As an admittedly overly simplistic example, a lexon mined from a relational table $R(A_1, \ldots, A_n)$ could be $<\gamma\ R\ r\ A_i>$ where γ is the application context and r is a suitable role played by attribute A in table R. Interesting research issues about context levels within ontologies arise here as one e.g. needs to separate local jargon from "common knowledge". Other important sources are numerous existing thesauri and glossaries, for instance the elaborate SAP® Glossary (for the crucial business process domain) in which each entry however needs to be individually analyzed to extract its knowledge structure (within DOGMA this is currently attempted experimentally using a version of ORM [Halpin, 1996]). The advantages for a more comprehensive and consistent corporate knowledge management using such ontobases should, however, already be quite obvious in spite of their simple basic structure and organization.

Finally, it is perhaps enlightening to see how ontologies in a sense may achieve a form of "semantics independence" for information- and knowledge based systems: just as database schemas achieved data independence by making the specification and management of stored data elements external to application programs, ontologies now will allow to specify and manage domain semantics external to those programs as well. Exactly how much knowledge is representable externally in this way will depend of course on the extent of the ontobase and on the manner constraints, rules, and application code make use of these knowledge elements. Conceivably at one point it will become economical to enforce the building of information systems, especially those destined for internet use or interoperation, by prescribing the use of *controlled vocabularies* which map explicitly to ontologies. Such vocabularies (including their rules of semantically correct usage) may even become a strategic resource for an organization, e.g. as part of a repository for corporate knowledge management.

4. http://www.starlab.vub.ac.be

4 Intelligent Information Access

- Carole Goble: Exploiting ontology reasoning services for web retrieval

  The web is increasingly viewed as a database, a knowledge base or a document collection. However, the web's original model was that of a hypertext. The notion of a hypertext includes navigation between resources as well as searching and resource discovery.

  Various generations of hypertext research have moved from static, embedded links manually crafted with rather poor semantics to their meaning, to the notion of Conceptual Hypermedia, where metadata descriptors are used to describe resources, schemas describe the contents of resources, and associations (or links) are derived through querying the schema. An ontology can take on the role of such a schema, effectively indexing the descriptors attached to the resources and navigating those resources by dint of navigation through the ontology.

  I suggest in the talk that ontology-based retrieval can benefit from reasoning services, specifically those services offered by a Description Logic, in a number of both indirect and direct ways. Indirectly, automated classification, subsumption testing and coherency satisfaction testing support the ontology development process and assist in the formation and management of metadata descriptors. Directly, we can again use the reasoning services to support classification-based retrieval and reasoning about query descriptors.

  In this talk I remind us that the web has a role as a hypertext, and present the past work on conceptual hypermedia. I argue that not only should ontologies support resource searching through their role as controlled vocabularies but also they can be used to support navigation in a conceptual hypertext. I then discuss two projects: STARCH, which uses an ontology implemented in a Description Logic for classification-based retrieval of stock photography images; and COHSE, which links an ontology service implemented using a Description Logic with an open hypermedia framework in order to experiment with ontology-based hypermedia navigation. Both actively seek to exploit the DL reasoning services in the deployment of the ontology. The talk concludes by raising a number of questions: what is the difference between searching and navigation? Are querying and link following the same? and What impact would an ontology have on navigation as opposed to resource discovery? Further, context and rhetoric are important issues in hypertext -- are these issues still important in the Semantic Web?

- Keith van Rijsbergen: What can IR offer the Web?

  Research in IR has a long and chequered history. Some trace it back to the original work of Vannevar Bush, others to the earlier work of Robert Fairthorne. The accumulated IR knowledge thus spans a sixty to seventy year period of research. One of the factors responsible for the strenght of the subject is its well developed and well founded experimental methodology which has made it easier for theory to affect practice. I will have some things to say about this by way of introduction. During the abovementioned period a number of significant research strands have emerged which continue to generate new and exciting work. I will present some of these strands or "dimensions", highlighting achievements and problems. Among the dimensions I will cover are, matching, inference, retrieval models, query defintion, language models, logics, etc. At each stage I will attempt to give an indication of the state-of-the art, and I will attempt to indicate whether these peculiarly IR research interests have any bearing on designing and building more effective retrieval tools

for WWW.

- Mounia Lalmas: An integrated solution for searching broadcast and web data in SAMBITS.

  The advent of digital TV creates the potential for convergence between conventional, broadcast content and narrowcast, internet based, content. Broadcasters can deliver greater depth of coverage through web sites that are tightly integrated with programmes and exploit the greater levels of interactivity and audience/user involvement that these technologies make available. Users can take greater control of the what and when they view and can take advantage of the resources of the wider internet that convergence will make available.

  Our aim is to develop a real-time consumer-type terminal prototype, which allows the demonstration and evaluation of integrated digital video broadcast and internet services, including local and remote interactivity. The terminal will provide integrated access to high quality digital video, as provided by DVB, and to the vast worldwide collection of interactive services and databases on the Internet. MPEG7 will provide the means for describing (MPEG2 and MPEG4) content with metadata. MPEG7 will thus add the functionality to filter transmitted multimedia content automatically and to search for multimedia as well as internet content on request according to the users profile and preferences.

  We will provide a focussed search of MPEG7 and web data to ensure minimum cognitive overload. Studies in information retrieval show that combining querying and browsing accesses to information helps users finding what they are looking for. The search engine will be based on HySpirit, an experimental platform at QMW for investigating the indexing and retrieval of information. HySpirit allows the representation of fact, content, and structure, and is therefore well suited to manipulate MPEG7 and web data.

- Austin Tate: Task Achieving Agents on the Web

  An important class of problems is related to activity. The "doing of things" is at the heart of human endeavour. The WWW has primarily concentrated to date on information storage and retrieval of information and other material. The data models and standards mostly relate to such things. I would like to see an emphasis placed on modelling activity and the collaboration between human and system agents that can be conducted through the WWW.

  The AI and process modelling community have started to develop shared models and ontologies to represent activities, tasks, agent capabilities, constraints, etc. These might form a generic core shared ontology to support the movement of information about activities over the WWW.

  The talk describes some work on producing collaborative, multi-agent systems with a mix of human and system agents engaging in planning and plan execution support over the WWW. The work includes O-Plan, I-X and the <I-N-CA> ontology for activity.

- Wolfgang Wahlster: Generating Virtual Web pages

  We introduce the concept of a virtual webpage and discuss the role of high-level ontological annotations for the generation of such third generation autoadaptive webpages. A virtual webpages is generated on the fly as a combination of various media objects from multiple web sites or as transformation of a real webpage. It looks like a real webpage, but is not persistently stored. A virtual webpage integrates generated and retrieved material in a coordinated way. It can be tailored to a user profile and adapted to a particular interaction context. It has an underlying representation of the presentation context so that an interface

agent can comment, point to, and explain its components.

We show how information extraction agents exploit ontologies during the information gathering process for virtual webpages. The plan-based approach to the generation of virtual webpages is presented and its use of a special mark-up layer is discussed. The presentation planner can generate virtual webpages for SMIL, WML, PML (Persona Mark-up Language) or the MS agent controller. We show how virtual webpages can be enabled for the interaction with life-like presentation agents. The plan operators provide high-level specifications of temporal and spatial design constraints, so that an autoanimated presentation agent is automatically synchronized with the dynamic elements of the virtual webpages.

We conclude that the generation of virtual webpages is heavily based on ontological annotations and that these ontologies are needed not only for information extraction agents but also for presentation agents.

- Henry Lieberman: Static vs. Dynamic Semantics of the Web

  Some of the semantics of the Web is determined by its static structure, if we view the Web as a relatively static database of linked HTML pages. But some of the semantics of the Web is also dynamic. A growing number of Web sites have dynamic content, or frequently update content. Exploration of Web sites by users is a dynamic process, and users have individual needs and desires that need to be taken into account by tools that provide personalized views of Web sites. I argue that the Web has not been well served by the old query-and-retrieval models that come from the field of traditional information retrieval. We need to view Web browsing as a dynamic, real-time activity, an exploration process that takes place cooperatively between between clients and servers, between interactive users and automated agents.

- Richard V. Benjamins: IBROW in the context of the semantic web

  The main objective of IBROW is to develop an intelligent brokering service able to retrieve knowledge components from distributed digital libraries, according to stated user requirements. The services will go beyond simple component retrieval and will include dynamic configuration of distributed, heterogeneous applications out of pre-existing components retrieved from different libraries. The components concerned are problem-solving methods (generic algorithms) and ontologies. This service will provide software-controlled access to a wide range of distributed and heterogeneous digital libraries of reusable knowledge components, at a level which abstracts from the underlying technology. In the envisaged scenario digital libraries are viewed as active, competence-based components that encapsulate reasoning services, such as configurable information filters, automatic classifiers and design problem solvers.

- Yolanda Gil: Knowledge Mobility: Semantics for the Web as a White Knight for Knowledge-Based Systems

  One of the challenges for knowledge-based systems is interoperation with other systems, intelligent or not. In recent years, my research group has participated in various such integration efforts, where interoperation was supported through translation techniques, mostly at the syntactic level and occasionally supported through ontology-based approaches. In this talk, I will argue that the interoperation challenge cannot be met with current approaches, since they entail trapping knowledge into formal representations that are seldom shareable and often hard to translate. I will propose an approach to develop knowledge bases that

captures at different levels of formality and specificity how each piece of knowledge in the system was derived from original sources, which are often Web sources. If a knowledge base contains a trace of information about how each piece of knowledge was defined, it will be possible to develop interoperation tools that take advantage of this information. The contents of knowledge bases will be more mobile and no longer be confined within a formalism. The Semantic Web will provide an ideal framework for developing knowledge bases in this fashion. We are planning to investigate these issues in the context of TRELLIS, a newly funded project motivated by military intelligence analysis. Starting from raw information sources, most of them originating on the Web, users will be able to add connections between selected portions of those sources. These connections may be initially very high level and informal, and the system will help users to formalize them further when other users request so or when they need to be related to other connections.

5 Applications and Demos

- Monica Crubezy: Protégé-2000

  Protégé-2000 is an integrated software tool used by system developers and domain experts to develop knowledge-based systems. Applications developed with Protégé-2000 are used in problem-solving and decision-making in a particular domain.

  The Protégé-2000 tool accesses all of these parts through a uniform GUI (graphical user interface) whose top-level consists of overlapping tabs for compact presentation of the parts and for convenient co-editing between them. This "tabbed" top-level design permits an integration of (1) the modeling of an ontology of classes describing a particular subject, (2) the creation of a knowledge-acquisition tool for collecting knowledge, (3) the entering of specific instances of data and creation of a knowledge base, and (4) the execution of applications. The ontology defines the set of concepts and their relationships. The knowledge-acquisition tool is designed to be domain-specific, allowing domain experts to easily and naturally enter their knowledge of the area. The resulting knowledge base can then be used with a problem-solving method to answer questions and solve problems regarding the domain. Finally, an application is the end product created when the knowledge base is used in solving an end-user problem employing appropriate problem-solving, expert-system, or decision-support methods.

- Jürgen Angele, Rudi Studer: Semantic Community Web Portals - KA2: The Community Web Portal of the Knowledge Acquisition Community

  Community web portals serve as portals for the information needs of particular communities on the web. The demo shows how a comprehensive and flexible strategy for building and maintaining a high-value community web portal has been conceived and implemented. The strategy includes collaborative information provisioning by the community members. It is based on an ontology as a semantic backbone for accessing information on the portal, for contributing information, as well as for developing and maintaining the portal. We have also implemented a set of ontology-based tools that have facilitated the construction of our show case - the community web portal of the knowledge acquisition community.

- Robert Jasper: Enabling Task Centered Knowledge Access through Semantic Markup

  Both commercial software companies and researchers have spent an enormous amount of effort developing ways to help users “find the right piece of information”. For the most part, this has been done without recognition of the user’s broader intentions or goals (i.e., the tasks they are performing). We believe that explicit recognition, representation, and exploitation of knowledge about the user’s goals and tasks they are performing is critical to exploiting the wealth of knowledge on the WWW.

  We describe an industrial application, which requires the user to locate and apply a number of resources for a variety of tasks supporting an airline helpdesk. Often, users are left on their own to determine whether, when, and how to navigate through a series of interfaces to support a given task. We’ve developed a centralized approach that takes into account important contextual information regarding users and the tasks they are performing.

  This approach leverages semantic markup to categorize and describe a variety of resources and their properties. We describe a powerful and flexible mechanism for dynamically constructing web interfaces tailored to a particular user and task. To accomplish this we embed semantic queries in XML templates, which return information about relevant resources. We are currently using RDF, F-Logic and the SiLRI query server. However, different query languages and servers may also be supported and interleaved in a single HTML template.

- Enrico Motta: Enabling knowledge creation and sharing on the World-Wide-Web

  The World-Wide-Web has traditionally been seen as a large hypertextual structure. However, recent developments in mark-up languages have introduced new perspectives: the web as a structured database or the web as a semantic knowledge base. In this talk I have introduced a fourth perspective, which I referred to as the knowledge web. That is, the web as the locus in which knowledge is created and shared. Within this perspective I have discussed a number of web-based technologies developed at the Knowledge Media Institute of The Open University in UK, which support knowledge creation and sharing over the web. These technologies include tools supporting document-centred discussion and debate, tools for collaborative ontology development, high-level interfaces supporting semantic queries and publishing tools. In the talk I also stressed the importance of a holistic approach to knowledge creation and sharing on the web, which takes into account a number of organizational, technological and user-centred issues, to ensure the feasibility of the proposed solutions. These ideas were illustrated with examples taken from a variety of domains, including guideline-centred healthcare, digital libraries and electronic publishing.

- Nicola Guarino: Onto Seek Approach for Ontology-driven Access to the Web

  Current information-retrieval techniques either rely on an encoding process using a certain perspective or classification scheme to describe a given item, or perform a full-text analysis, searching for user-specified words. Neither case guarantees content matching, because an encoded description might reflect only part of the content, and the mere occurrence of a word (or even a sentence) does not necessarily reflect the document's content. For general documents, there doesn't yet seem to be a much better option than some sort of lazy full-text analysis, leaving us to sift through endless results pages. However, if we narrow the field to a relevant class of information repositories‹online yellow pages and product catalogs‹content-retrieval techniques based on simple representation capabilities and large linguistic ontologies can be both feasible and crucial. We developed OntoSeek, our information-retrieval system,

to target these repositories. In this article, we discuss the special characteristics of online yellow pages and product catalogs, examine linguistic ontologies[1] role in content matching, and present OntoSeek´s architecture.

- Ian Harrocks: "OIL"

  Exploiting the full potential of the World Wide Web will require semantic as well as syntactic interoperability. This can best be achieved by providing a further representation and inference layer that builds on existing and proposed web standards. The OIL language extends the RDF schema standard to provide just such a layer. It combines the most attractive features of frame based languages with the expressive power, formal rigour and reasoning services of a very expressive description logic.

  Reasoning with OIL ontologies can be achieved via a translation to a semantically equivalent terminology in the SHIQ description logic, for which a sound and complete yet highly efficient reasoning engine has already been implemented in the FaCT system. FaCT's CORBA based client server architecture means that it can easily be integrated with tools and applications in order to provide them with reasoning services. This was illustrated by demonstrating an intelligent CASE tool that uses FaCT's reasoning services to support schema integration and validation. More details about OIL can be found at:

  http://www.ontoknowledge.org/oil/oilhome.shtml

  FaCT is freely available from:

  http://www.cs.man.ac.uk/fact

- Jos van der Meer and Frank van Harmelen: AIdministrator Nederland: The AIdministrator Information Map Generator: overviews of large collections of information

  Easy accessibility of huge amounts of information (for instance on Intranets) becomes an ever more important condition for effective knowledge management. High demands are made upon disclosure of such information: not only must the information be always up-to-date and available, but it must also be classified in a meaningful way and be easily searchable. A related problem is that different persons have different information needs and therefore demand different kinds of classification and navigation structures.

  Existing search- and navigation-tools do not satisfy these high knowledge management demands. Search-tools are usually keyword based, resulting in a lot of undesirable information. The navigation structure is limited to hand-made menu's and index pages and must, once it has been created, be used by everyone. One additional disadvantage is that this way of structuring quickly ages and therefore requires a lot of maintenance.

  The AIdministrator Information Map Generator does provide the means to meet the high requirements of knowledge management. The Information Map Generator can generate semantically organized information maps of document collections (web sites, Intranets). These information maps are graphical overviews based on the contents of the documents, whereas the documents are grouped based on freely definable categories.

  • Step 1 (human): The builder or administrator of a site uses the AIdministrator Information Map Generator to describe the different content categories for the site.

  • Step 2 (automatic): The AIdministrator Information Map Generator automatically classifies all pages according to these content categories.

• Step 3 (automatic): The AIdministrator Site Map Generator automatically generates a visual map based on the content categories.

- Deborah McGuinness: The Chimaera Ontology Environment

  Large-scale ontologies are becoming an essential component of many applications including standard search (such as Yahoo and Lycos), e-commerce (such as Amazon and eBay), configuration (such as Dell and PC-Order), and government intelligence (such as DARPA's High Performance Knowledge Base (HPKB) program). The ontologies are becoming so large that it is not uncommon for distributed teams of people with broad ranges of training to be in charge of the ontology development, design, and maintenance. Standard ontologies (such as UNSPSC) are emerging as well which need to be integrated into large application ontologies, sometimes by people who do not have much training in knowledge representation. This process has generated needs for tools that support broad ranges of users in (1) merging of ontological terms from varied sources, (2) diagnosis of coverage and correctness of ontologies, and (3) maintaining ontologies over time. In this demonstration, we present a new merging and diagnostic ontology environment called Chimaera, which was developed to address these issues in the context of HPKB.

6 Working Groups

- User scenarios, applications, & evaluation (chair: Henry Lieberman and Mounia Lalmas)

  What can the "semantic Web" be used for? While much of the seminar concerned how to represent knowledge on the Web, this workshop focused on what use could be made of that knowledge, specific user scenarios, user interface and methods for evaluating semantic Web applications. We started out by having trouble hooking up the computer projector, which let to a discussion of how a semantic Web might assist the user for this scenario -- call up the manufacturer for instructions, automated negotation between a laptop and its projector, etc. Travel planning was another domain discussed -- where up-to-the-minute information and figuring out what to do when things go wrong are important. In games, amazingly enough, if you consider how much time people put into constructing Mud or Doom worlds, people actually do semantic markup for fun! The semantic Web could facilitate agents that provide context-sensitivity, personalization, or make use of history. People could express their goals in a more high level form. Finally, though we didn't explore this issue enough, we considered how to evaluate such applications. We noted that as some semantic applications are now being added to more conventional tools [as the Cyc system is being used to augment a conventional search engine], direct comparison of semantically-enabled and conventional applications becomes possible.

- Semantic interoperability and top-level ontologies (chair: Jérôme Euzenat, Nicola Guarino)

  In the context of the so-called semantic web, knowledge expressed in a formal way will be disseminated through the web. This aims at facilitating the understanding of the content of documents, especially by machines. In order to take advantage of this valuable knowledge, application will gather knowledge pieces from the web. They will have to interpret the various knowledge sources in a coherent way if they want to carry on a satisfaying reasoning.

Several tools for assisted ontology merging have been demonstrated during the seminar.

However, it is not clear that interoperability between these models (knowledge bases or ontologies) will be easily achieved. There are at least two levels in which a "semantic mismatch" can occur resulting in a failure to interoperate:

• When different knowledge representation languages are used, as it might be the case, there is a need for understanding the semantics of the languages involved in order to perform importation of model fragments. This can be achieved by active translators able to expand the representations that cannot be literally understood by the target language or by the description of the formal semantics of languages in order to check properties (completeness, correctness...) against translations.

• Even when the same representation language is used, it is not straightforward to preserve the intended meaning of a concept partially described in that language. Top-level ontologies can play the role of anchoring one ontology with regard to the other and reducing the set of possible interpretations of the concepts manipulated.

When setting up a future semantic web, it is important to be aware of these problems that occur even in the simple cases.

- Stefan Decker: "Specific Requirements for KR on the Web"

  The Web is a unique chance for the Knowledge Representation (KR) Community, since never before was a infrastructure available, that enables knowledge interchange on the scale of the Internet.

  Knowledge is produced by every user on the web - often for free and once a Knowledge source is established it is usually public available. Standardization efforts are going on to provide the basis for KR techniques on the Internet, and since B2B e-commerce as well as service integration efforts enforce shared semantics and declarative specifications, also the interest and the willingness to invest in KR topics is very high. Perfect solutions are not required - often a partial solution already enables new applications. However, also new challenges for the KR community arise:

  • Since the Web is large and still growing fast, scalability of individual Knowledge Representation Techniques has to be ensured.

  • The Web has many different authors and no central authority, hence correctness and trustworthiness of the available Knowledge can not be guaranteed. So mechanisms to establish trust between knowledge sources need to be established. Also the content itself is highly diversified: a single KR technique is unlikely to succeed, thus a variety of KR techniques is needed to represent capture the available knowledge.

  • These KR techniques are available, however, interoperability among them is hard to establish. Interoperability is necessary on the Web, since the "network effect" can only occur if the knowledge can be used and reused by many different services and people. This aspect is especially significant, since the usage scenarios for the represented knowledge can not be predicted. Hence the knowledge has to be defined using principles, that permit as much task independence as possible.

  The Web presents new research challenges to the KR Community, but also a unique chance to put ideas of KR and AI into practical use.

7 Conclusions

Tim Berners-Lee, director of the World Wide Web consortium, has referred to the future of the current WWW as the "semantic web" -- an extended web of machine-readable information and automated services going way beyond current capabilities. The explicit representation of the semantics underlying data, programs, pages, and other web resources, will enable a knowledge-based web providing a qualitatively new level of service. Automated services will be better able to assist humans in achieving their goals by "understanding" more of the content on the web and thus providing more accurate filtering, categorization, and search of information sources. This process will ultimately lead to an extremely knowledgeable system with various specialized reasoning services that will support us in nearly all aspects of our daily life -- making access to information as pervasive, and necessary, as access to electricity is today.

In the systems of the future, information will not be simply a set of passive entities residing in a repository. Instead, active information sources will play a critical role accessed via network-enabled, information-provision services. These services will not only support better extraction and search, but will also more directly support human task achievement. To make this possible, machine-understandable representation of semantics is required for the automated selection and combination of these reasoning services.

A key enabler for the semantic web is on-line ontological support for data, information and knowledge exchange. Given the exponential growth of on-line information available, automatic processing becomes mandatory for keeping it managed and accessible. Being used to describe the structure and semantics of information exchange, ontologies will play a key role in areas such as knowledge management, B2B e-commerce and other such burgeoning electronic enterprises.

In this workshop, many cutting edge papers were presented describing the state of the art in this emerging new research area. Participants from a number of different organizations described research activities in academia, industry and government. In addition, a description of the US Defense Advanced Research Project Agency's DAML (DARPA Agent Markup Language) project, was presented, and possible EC funding plans for this work were discussed. Many follow-on activities to the Dagstuhl seminar are now being planned, including: setting up a scientific journal as part of the Electronic Transactions on AI, submitting IST proposal on research projects and thematic networks, and organizing follow-up Transatlantic workshops on web-semantics efforts.

More information on current and future efforts can be found at www.ontoknowledge.org, www.semanticweb.org, and www.daml.org.

Acknowledgment. We all would like to thank the staff of Dagstuhl and Rainer Faulstich for their excellent support. Last but not least, we should not forget to thank Nigel Shadbolt for his remarkable dinner speech.

References

[Genesereth & Nilsson, 1987]
M. Genesereth and N. Nilsson: *Logical Foundations of Artificial Intelligence*, Morgan Kaufman, 1987.
[ISO TR 9003]

ISO TR 9003: Concepts and Terminology of the Conceptual Schema and the Information Base, ISO Technical Report, International Standards Organization, 1990.

[Guarino, 1998]

N. Guarino: Formal Ontology and Information Systems. In *Proceedings of FOIS'98,* N. Guarino (ed.), IOS Press, 1998.

[Halpin, 1996]

T. Halpin: *Conceptual Schema and Relational Database Design*, 2nd Ed., Prentice Hall, 1996.

[Lassila 1998]

Ora Lassila: Web Metadata - a Matter of Semantics, *IEEE Internet Computing,* 2(4) 30-37, 1998.

[Meersman, 1999]

R. Meersman: Semantic Ontology Tools in Information Systems Design. In *Proceedings of the ISMIS'99 Conference,* Z. Ras and M. Zemankova (eds.), LNCS, Springer Verlag, 1999.

[Reiter, 1988]

R. Reiter: Towards a Logical Reconstruction of Relational Database Theory. In J. Mylopoulos and M.L. Brodie (eds.), *Readings in AI and Databases*, Morgan Kaufman, 1988.