1 The Future of the Internet

1.1 Introduction

In the beginning of the Internet days, software programmers developed all Web ages. Today, the Web provides perhaps the simplest way to share information, and literally everyone writes Web pages, with the help of authoring tools, and a large number of organizations disseminate data coded in Web pages. The *Hypertext Markup Language* (HTML) is typically the language used to code information about renderization (font size, color, position on screen, etc.) and hyperlinks to other Web pages or resources on the Web (multimedia files, text, e-mail addresses, etc.). The net result is that the Web keeps growing at an astounding pace, now having over eight billion Web pages. However, most Web pages are still designed for human consumption and cannot be processed by machines. Computers are used only to display the information, that is, to decode the color schema, headers, and links encoded in Web pages.

Furthermore, Web search engines, the most popular tools to help retrieve Web pages, do not offer support to interpret the results. For that, human intervention is still required. This situation is progressively getting worse as the size of search results is becoming too large. Most users only browse through the top results, discarding the remaining ones. Some search engines are resorting to artifice to help control the situation, such as indexing the search result, or limiting the search space to a relevant subset of the Web (such as in Google Scholar).

The conclusion is that the size of search results is often just too big for humans to interpret, and finding relevant information on the Web is not as easy as we would desire.

1.2 The Syntactic Web

Today's Web may be defined as the *Syntactic Web*, where information presentation is carried out by computers, and the interpretation and identification of relevant information is delegated to human beings. Of course, the interpretation process is very demanding and requires great effort to evaluate, classify, and select relevant information. Because the volume of available digital data is growing at an exponential rate, it is becoming virtually impossible for human beings to manage the complexity and volume of the available information. This phenomenon, often referred to as *information overload*, poses a serious threat to the very usefulness of today's Web. The question is: Why can't computers do this job for us?

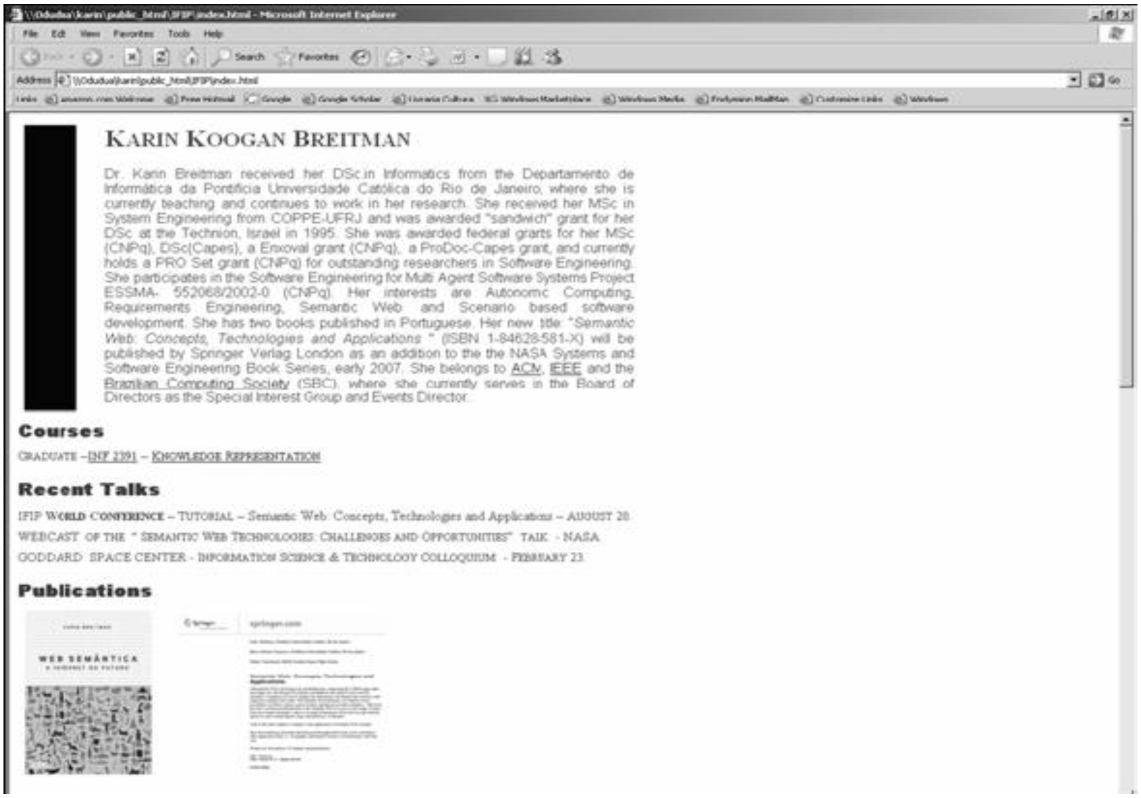
One of the reasons resides in the fact that Web pages do not contain information about themselves, that is, about their contents and the subjects to which they refer. We can make an analogy with a library where books, instead of being organized by subject, are randomly displayed. Every time we wanted to borrow a book, we would have to search for it based on title and related words. Imagine that we wanted to learn about the TCP/IP protocol. We would have to look for a book about networks. If we only used "network" as a keyword, we would retrieve computer science books, as well as books about telephone and electrical networks. We would then be responsible for filtering and selecting those books that are of genuine interest. This is precisely the situation we are dealing with in the Syntactic Web of today.

Web search engines do help identify relevant Web pages, but they suffer from the following limitations.

- Search results might contain a large number of entries, but they might have low recall precision. For example, a search for Web pages where “TCP/IP” and “protocol” occur might return all relevant Web pages, but the result would be of very little use if the user had to sift through 39,857 Web pages of little interest.
- Search results are sensitive to the vocabulary used. Indeed, users frequently formulate their search in a vocabulary different from that which the relevant Web pages adopt. In the TCP/IP example, the relevant Web pages might use “standard”, instead of “protocol”; hence, these Web pages would not be the best match for a search using the keywords “TCP/IP” and “protocol”.
- Search results appear as a list of references to individual Web pages. However, it is often the case that, among the Web pages listed in the search result, there are many entries that belong to the same Web site. Conversely, if the relevant information is scattered in more than one entry, it is difficult to determine the complete set of relevant entries.

The conclusion is that the Web has evolved as a medium for information exchange among people, rather than machines. As a consequence, the semantic content, that is, the meaning of the information in a Web page, is coded in a way that is accessible to human beings alone. Figures 1.1 and 1.2 illustrate the difference between how humans and computers perceive a Web page, dramatizing why we urgently need to add more semantics to the Web pages so they can be processed by machines as well as humans.

1.2 The Syntactic Web 5


\\Odudua\karin\public_html\IFIP\index.html - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Search Favorites
Address \\Odudua\karin\public_html\IFIP\index.html
Go
Links

KARIN KOOGAN BREITMAN

Dr. Karin Breitman received her DSc in Informatics from the Departamento de Informática da Pontifícia Universidade Católica do Rio de Janeiro, where she is currently teaching and continues to work in her research. She received her MSc in System Engineering from COPPE-UFRJ and was awarded "sandwich" grant for her DSc at the Technion, Israel in 1995. She was awarded federal grants for her MSc (CNPq), DSc(Capes), a Enxoval grant (CNPq), a ProDoc-Capes grant, and currently holds a PRO Set grant (CNPq) for outstanding researchers in Software Engineering. She participates in the Software Engineering for Multi Agent Software Systems Project ESSMA- 552068/2002-0 (CNPq). Her interests are Autonomic Computing, Requirements Engineering, Semantic Web and Scenario based software development. She has two books published in Portuguese. Her new title: "*Semantic Web: Concepts, Technologies and Applications* " (ISBN 1-84628-581-X) will be published by Springer Verlag London as an addition to the the NASA Systems and Software Engineering Book Series, early 2007. She belongs to ACM, IEEE and the Brazilian Computing Society (SBC), where she currently serves in the Board of Directors as the Special Interest Group and Events Director.

Courses

GRADUATE –INF 2391 – KNOWLEDGE REPRESENTATION

Recent Talks

IFIP WORLD CONFERENCE – TUTORIAL – Semantic Web: Concepts, Technologies and Applications – AUGUST 20

WEBCAST OF THE " SEMANTIC WEB TECHNOLOGIES: CHALLENGES AND OPPORTUNITIES" TALK - NASA GODDARD SPACE CENTER - INFORMATION SCIENCE & TECHNOLOGY COLLOQUIUM - FEBRUARY 23

Publications

Figure 1.1 how humans see a Web page

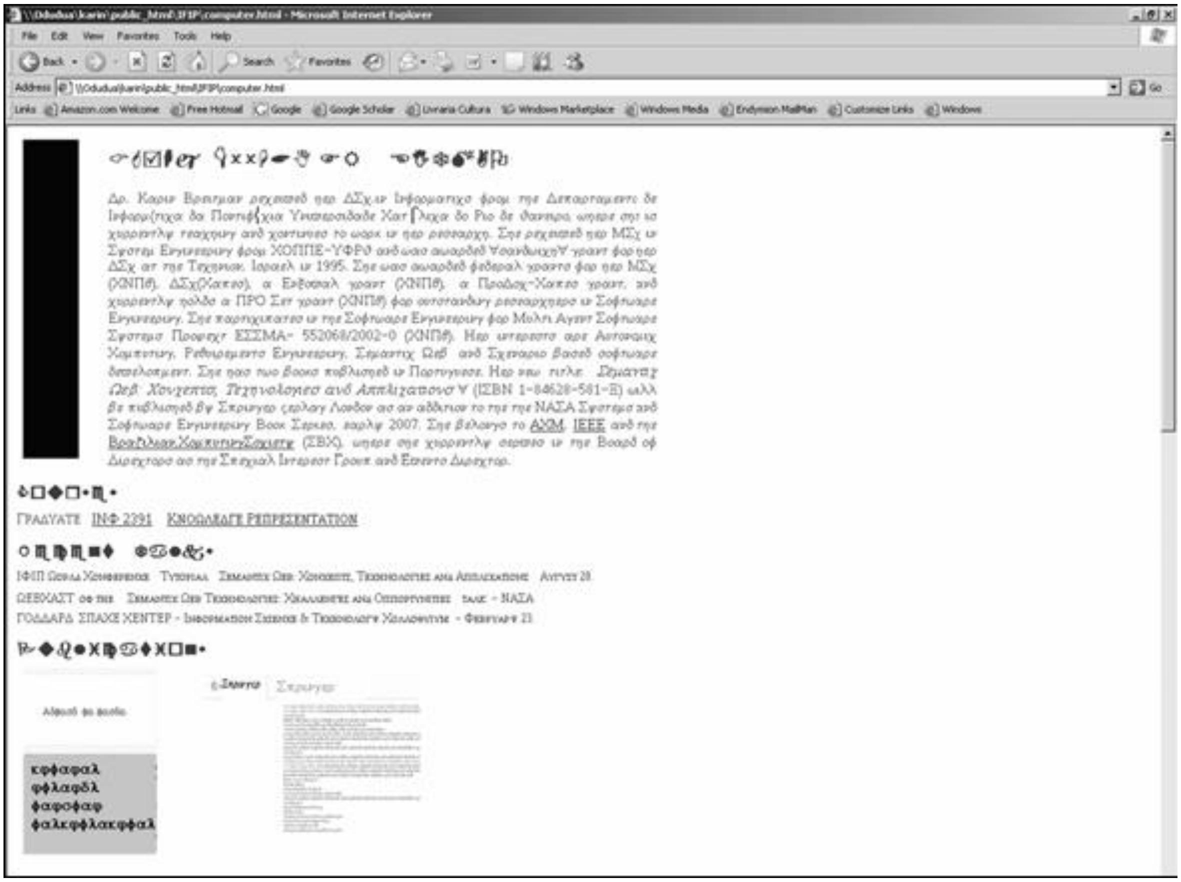

\\Odudua\karin\public_html\IFIP\computer.html - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Address \\Odudua\karin\public_html\IFIP\computer.html
Go
Links Amazon.com Welcome Free Hotmail Google Google Scholar Livraria Cultura Windows Marketplace Windows Media Endymion MailMan Customize Links Windows

Figure. 1.2 How computers see the same Web page

1.3 The Semantic Web

In 2001, Berners-Lee, Hendler, and Lassila published a revolutionary article in the magazine *Scientific American*, entitled “The Semantic Web: A New Form of Web Content That Is Meaningful to Computers Will Unleash a Revolution of New Possibilities” (Berners-Lee et al. 2001). In this article, the authors describe future scenarios in which the Semantic Web will have a fundamental role in the day-to-day life of individuals. In one of the scenarios, Lucy needs to schedule a series of medical consultations for her mother. A series of restrictions applies to this scenario: Lucy’s tight schedule, geographical location constraints, doctor’s qualifications, and adherence to their Social Security plan.

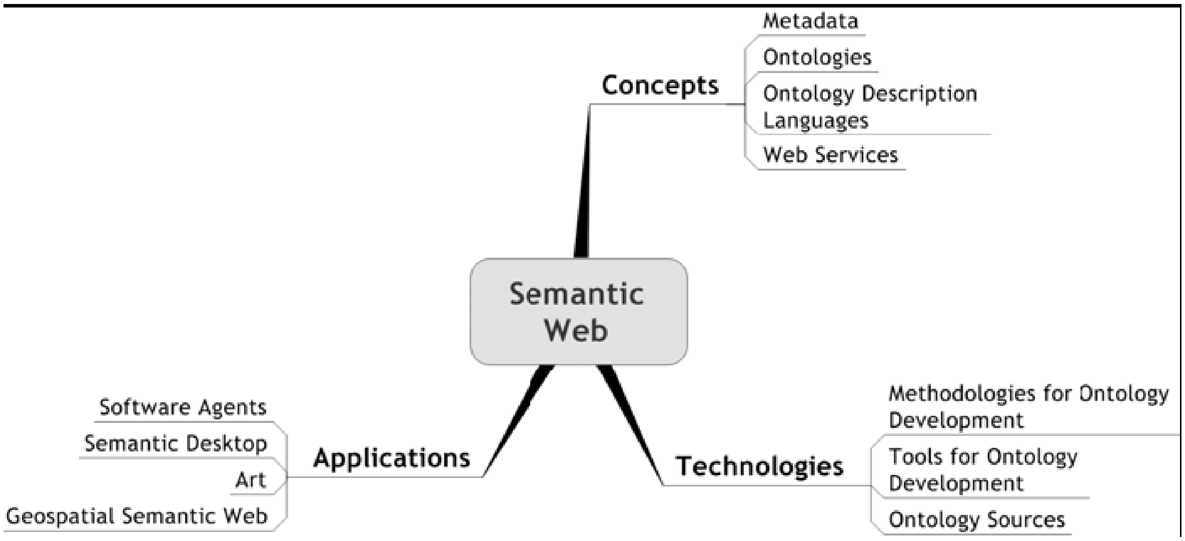
To help Lucy find a solution, there is a software agent, capable of negotiating among different parties: the doctor, Lucy’s agenda and medical services directory, among others. The point is that, although each party codes its information in a different way, because of a semantic layer, they are able to interact and exchange data in a meaningful way. The enabling technology that will bring this scenario forward is what the authors called the Semantic Web. The authors emphasized the important point that most of the actions described in the scenarios can be achieved in the Syntactic Web of today, but not without considerable effort and many comes-and-goes between different Web sites. The promise of the Semantic Web is that it will unburden users from cumbersome and time-consuming tasks.

1.4 How the Semantic Web Will Work

In order to organize Web content, artificial intelligence researchers proposed a series of conceptual models. The central idea is to categorize information in a standard way, facilitating its access. This idea is similar to the solution used to classify living beings. Biologists use a well-defined taxonomy, the Linnaean taxonomy, adopted and shared by most of the scientific community worldwide. Likewise, computer scientists are looking for a similar model to help structure Web content.

On the other hand, it is believed that the huge success of the Web is due to the freedom it provides. In the same environment, we can find very sophisticated Web sites, designed by specialists, and personal Web pages, created by individuals with little or no computer expertise. There is no censorship to the quality of the information in a Web page either; it virtually depends on the Web page owner's discretion. In the Web, scholarly papers cohabit peacefully with vendors' Web sites and personal blogs. In this scenario, anarchical at best, it is very hard to imagine that a single organization model could prevail.

Similarly to the Syntactical Web, the Semantic Web should be as decentralized as possible, asserts Berners-Lee (Berners-Lee et al. 2001). However, the fact that there should be no central control requires many compromises, the most important being to give up the consistency ideal. Hendler, from the University of Maryland and one of the founding fathers of the Semantic Web, believes that, in the future, instead of a single information organization model, we will have a series of parallel models (Hendler 2001). Hendler's prediction is that every business, enterprise, university, and organization on the Web of the future will have its own organizational model or ontology.



How will this "Web of the Future" be effectively built, no one really knows. Although guesses vary from author to author, some themes are recurrent in most discussions. We illustrate the most important ones in Fig. 1.3 and discuss them separately in the remainder of this section.

Metadata

Metadata are data about data. They serve to index Web pages and Web sites in the Semantic Web, allowing other computers to acknowledge what the Web page is about.

Knowledge organization dates back from antiquity. The Greek philosopher Aristotle provided the first known solution with his category system. He proposed that all knowledge should be structured in categories, organized under supertypes (genus) and subtypes (species). In Table 1.1, we illustrate the categories proposed by Aristotle.

Traditional use of metadata was often limited to a relatively few participating institutions, such as libraries and museums. The use of metadata was mostly restricted to the cataloguing of specific collections, such as works of art, which typically consisted of a limited, thus enumerable, number of physical objects.

The use of metadata in the context of the Semantic Web is somewhat similar, except for the fact that the number of institutions and the number of objects—Web pages—are orders of magnitude larger, and here lies the problem. Indeed, in the Semantic Web, we want to catalogue an enormous number of resources, mostly virtual, distributed all over the world, coded in different languages, by different groups.

Ontologies

The word ontology comes from the Greek *ontos* (being) + *logos* (word). It was introduced in philosophy in the nineteenth century by German philosophers to distinguish the study of being as such from the study of various kinds of beings in the natural sciences. As a philosophical discipline:

The subject of Ontology is the study of the categories of things that exist or may exist in some domain. The product of such a study, called an ontology, is a catalogue of the types of things that are assumed to exist in a domain of interest D from the perspective of a person who uses a language L for the purpose of talking about D. The types in the ontology represent the predicates, word senses, or concept and relation types of the language L when used to discuss topics in the domain D. (Sowa 1997)

In computer science, ontologies were adopted in artificial intelligence to facilitate knowledge sharing and reuse (Fensel 2001; Davies et al. 2003). Today, their use is becoming widespread in areas such as intelligent information integration, cooperative information systems, agent-based software engineering and electronic commerce.

Ontology is defined in Guarino (1998) as “an artifact, constituted by a specific vocabulary used to describe a certain reality, plus a set of explicit assumptions regarding the intended meaning of the vocabulary.” Ontologies are conceptual models that capture and make explicit the vocabulary used in semantic applications, thereby guaranteeing communication free of ambiguities. They will be the *lingua franca* of the Semantic Web.

In Chapter 2, we discuss ontologies, their origins, formalisms, types, and basic elements and, in Chapter 9, we present upper-level and domain ontologies, and list ontology libraries.

Formal Systems

Formal systems provide the ability to deduce new sentences from existing sentences using specific inference rules. This ability, referred to as logical inference, is an essential component of Semantic Web ontology formalism.

To ensure effective information sharing among software agents, ontologies will need to be expressive enough to establish a common terminology that guarantees consistent interpretation. Because first-order logic is known to be intractable, the Semantic Web community has been exploring the possibilities of adopting description logic as the paradigm formal system. Briefly, description logic models the application domain by defining the relevant concepts of the domain and then using these concepts to specify properties of objects and individuals occurring in the domain.

In Chapter 3, we provide a concise introduction to description logic.

Ontology Description Languages

Ontology description languages are specifically designed to define ontologies. They recently received considerable attention, boosted by the emergence of the Semantic Web. This new breed of ontology description languages is sometimes called lightweight ontology languages, Web-based ontology languages, or markup ontology languages.

The Resource Description Framework (RDF) is a general-purpose language for representing information about resources in the Web and, to some extent, a lightweight ontology language. The lack of expressiveness of RDF was partly eased with the introduction of the RDF Vocabulary Description Language 1.0: RDF Schema (RDF Schema or RDF-S), which offers primitives to model hierarchies of classes and properties.

The Ontology Inference Layer (Oil) is the result of the On- To-Knowledge Project, and has a formal semantics based on description logic. At about the same time Oil was developed, the Defense Advanced Research Projects Agency (DARPA) sponsored the DARPA Agent Markup Language (DAML) Program. These two languages were amalgamated into a single language, DAML+Oil. A reformatted version of DAML+Oil served as a starting point for the Web Ontology Language (OWL).

In Chapters 4, 5, and 6, we review RDF, OWL, and rule languages recently proposed for the Semantic Web.

Web Services

In the Semantic Web vision, the services provided by the Web of the future will be greatly extended and improved, if semantics is added to the present Web resources. Computers will be able to make doctor appointments, synchronized with our agenda, 10 1. The Future of the Internet find new suppliers for products we consume, and make traveling arrangements, among many other tasks.

In Chapter 7, we discuss the possibilities of semantic Web services, including OWL-S, an upper ontology to describe services.

Methodologies and Tools for Ontology Development

According to Jim Hendler, in the near future, most Web sites of interest will sport their own ontology (Hendler 2001). Therefore, the Web will be, quoting Hendler, a composition of a “great number of small ontological components consisting largely of pointers to each other.”

Tracing a parallel to the history of the Web itself, Hendler assumes that ontology creation will be conducted in the same, nearly anarchic and decentralized fashion as the more than eight billion Web pages have been created. The result will be a great variety of lightweight ontologies, created and maintained by independent parties.

His predictions seem to be true, as the number of tools for ontology edition, visualization, and verification grows. The best examples are the Protégé and OilEd tools, which sprung from large cooperation projects, involving many universities and different countries. With the increasing number of available books and online tutorials, crafting an ontology today is possibly no harder than creating a Web page was ten years ago. Our experience with ontologies has demonstrated that ontology development is not particularly challenging, compared to building other conceptual models used in our software engineering practice.

Evidently, the quality of the resulting ontology depends on the ability of the person engaged in the modeling, which is also true for most conceptual models. Indeed, the number of "lightweight" ontologies, that is, developed by independent groups and organizations rather than by knowledge engineers, is rapidly growing as can be verified by visiting some of the public ontology repositories, such as the DAML repository (URL: http://www.daml.org).

In Chapter 8, we discuss ontology development methodologies and, in Chapter 10, we put forward a discussion on what ontology development tools are presently available. With the help of examples, we discuss and explore the necessary building blocks, so that readers will be able to create their own Semantic Web ontologies.

Applications of Semantic Web Technologies

Applications for the concepts and technologies discussed in this book are by no means limited to indexing Web pages. Other areas provide excellent challenges and opportunities for such technologies, as discussed in the last part of the book.

For example, consider software agents, defined as autonomous software applications that act for the benefit of their users. According to Antoniou and Harmelen (2004), a personal agent in the Semantic Web will be responsible for understanding the desired tasks and user preferences, searching for information on available resources, communicating with other software agents, and comparing information so as to provide adequate answers to its users. Of course software agents will not be a substitute for people, who will ultimately be responsible for making the 1.5 What the Semantic Web Is Not 11 important decisions. To accomplish their tasks, software agents will make heavy use of metadata and ontologies in general.

Semantic desktop applications provide a second example from the area of software engineering. Such applications use ontologies to integrate desktop applications and the Web, facilitating personal information management, information distribution, and collaboration on the Web, beyond the mere sending of e-mails.

As a third example, now from a different area, we point out that cataloguing our cultural heritage has been a major activity of museums and other cultural institutions throughout the world. Today, almost all major museums make their collections available over the Web, often with remarkable quality, such as the Hermitage Museum Web site.

In parallel, many organizations have been working toward the development of metadata standards and controlled thesauri for describing cultural objects that facilitate their dissemination over the Web. Such efforts therefore contribute to and benefit from Semantic Web technologies.

A similar phenomenon occurs in the geospatial application area. We observe that the large volume of geospatial data available on the Web opens up unprecedented opportunities for data access and data interchange, facilitating the design of new geospatial applications and the redesign of traditional ones.

A convenient way to provide access to geospatial data over the Web is to implement Web services that encapsulate the data sources and that adopt Semantic Web technologies. This evolution includes the development and encoding of formal geospatial ontologies, which leverage existing standards. The result is called the Geospatial Semantic Web.

We discuss software agents and semantic desktop applications in Chapters 11 and 12, respectively. In Chapter 13, we address standardization efforts that combine metadata and controlled vocabularies to describe works of art. Finally, in Chapter14, we overview technologies that facilitate the development of the Geospatial Semantic Web.

1.5 What the Semantic Web Is Not

The Semantic Web Is Not Artificial Intelligence

The concept of machine-processable documents does not imply some sort of magic artificial intelligence (AI) that makes computers understand what humans mean. The concept means that computers will have enough semantics to allow them to solve well-defined problems through the sequential processing of operations. Instead of having computers that “understand” people’s language, we argue in favor of going the extra mile to make representations that are passive to automatic processing (that is, ontologies as opposed to free text files).

It is true that most techniques that are currently used in the Semantic Web did come from research in AI. Given a history of unsuccessful AI projects, it is reasonable to suppose that the Semantic Web may be fated for the same destiny. According to Antoniou and Harmelen (2004), this supposition is completely unfounded. The realization of the Semantic Web potential does not depend on some 12 1. The Future of the Internet sort of computational intelligence, as promised by some AI researchers (and most science fiction writers) some thirty years ago.

In the specific case of the Semantic Web, partial solutions are acceptable. It may be the case that a software agent does not come even close to conclusions a human being may be capable of, but still this software agent may contribute to building a better Web than that which we have today. Antoniou and Harmelen (2004) summarize this concept as follows:

"If the ultimate goal of AI is to build an intelligent agent exhibiting human-level intelligence (and higher), the goal of the Semantic Web is to assist human users in their day to day online activities."

The Semantic Web Is Not a Separate Web

The Semantic Web is not a separate Web, but rather an extension of the current Syntactic Web. In this new Web, information will have well-defined meaning as a result of the use of semantic markup languages. Such languages, essentially ontology description languages, will be added to existing Web pages, in an architecture called *The Semantic Web Wedding Cake Architecture* by Berners-Lee etal. (2001).

In Chapter 2, we explain this architecture in more detail.

The Semantic Web Will Not Demand the Use of Complex Expressions

Although the Semantic Web language standard, OWL, supports very sophisticated constructs, it will not be mandatory that every Semantic Web application shows this level of complexity. It is believed by the World Wide Web Consortium (W3C) that, for most Semantic Web Applications, the lighter species of OWL will be sufficient. Most applications that generate RDF markup will, in practice, be limited to simplified expressions, such as access control, privacy settings, and search criteria.

The Semantic Web Is Not a Rerun of a Failed Experiment

Another common question is the relationship between the Semantic Web and knowledge representation systems. Hasn't it all been tried before with the KIF (Knowledge Interchange Format) and CYC projects? The answer is not very direct as the goals of the two initiatives are different.

The goal of the knowledge representation community is to create canonical representations, that is, unique models that are to be used as reference, standards with which applications must comply. That is the case with large projects, such as CYC and SUMO, discussed in Chapter 9. The Semantic Web, on the other hand, is seeking the integration of different models. Such models, known as domain ontologies and predicted to proliferate in the near future, will be developed and maintained by independent parties, according to Hendler (2001). Of course, the experience gained in the construction of large knowledge representations, such as CYC, will be taken into consideration in the path to the Semantic Web (Hendler 2001).
References 13

1.6 What Will Be the Side Effects of the Semantic Web

We centered the discussion so far on the potential benefits of the Semantic Web, where computers will be able to understand the information available on the Web and take over tasks that users have been doing manually. Evidently, the goal of the Semantic Web is to create a Web that is more adequate for the users' needs. Emerging technologies will allow semantics to be added to existing Web pages and applications so as to make the Semantic Web vision come true.

Perhaps the most pervading benefit of the Semantic Web might come as a collateral effect of creating truly global knowledge representations. To this day, every area of human endeavor has been using proprietary conceptual models that are believed to best represent the specific knowledge permeating the area.

We have, for example, architectural plants, circuit designs, cartographic maps, artistic casts, object-oriented models, economy planning spreadsheets, and a myriad of other specific models. Specific conceptual models facilitate the coordination and communication among specialized communities. However, they make communication across different communities (or cultures) difficult. Even in the same field, we may find communities that use completely different conceptual models. This argues in favor of the adoption of ontologies, in the computer science sense.

The coordination of different communities (or subcultures), according to Berners-Lee, is “painfully slow and requires a lot of communication.” Of course, the use of a common language, a *lingua franca*, is essential to the process. Perhaps, in the context of the Semantic Web, such a language will emerge (Berners-Lee 2001).