

# Which Semantic Web?

Catherine C. Marshall

Microsoft Corporation  
1 Microsoft Way  
Redmond, WA 98052  
1 (425) 705 - 9057

cathymar@microsoft.com

Frank M. Shipman

Department of Computer Science  
Texas A&M University  
College Station, TX 77843-3112  
1 (979) 862 - 3216

shipman@cs.tamu.edu

## ABSTRACT

Through scenarios in the popular press and technical papers in the research literature, the promise of the Semantic Web has raised a number of different expectations. These expectations can be traced to three different perspectives on the Semantic Web. The Semantic Web is portrayed as: (1) a universal library, to be readily accessed and used by humans in a variety of information use contexts; (2) the backdrop for the work of computational agents completing sophisticated activities on behalf of their human counterparts; and (3) a method for federating particular knowledge bases and databases to perform anticipated tasks for humans and their agents. Each of these perspectives has both theoretical and pragmatic entailments, and a wealth of past experiences to guide and temper our expectations. In this paper, we examine all three perspectives from rhetorical, theoretical, and pragmatic viewpoints with an eye toward possible outcomes as Semantic Web efforts move forward.

## Categories and Subject Descriptors

H.5.4 [Information Interfaces and Presentation]:

Hypertext/Hypermedia – *architectures, theory, navigation*

## General Terms

Design, Experimentation, Standardization, Theory.

## Keywords

Semantic Web, Hypertext, Digital Libraries, Knowledge Representation, Knowledge Acquisition, Information Systems.

## 1. INTRODUCTION

Over the past decade, the Web has grown from what many perceived as an improved Gopher interface to become the new medium of communication. It would have been hard to predict such a transition; the doubts that many researchers had about this outcome turned out to be misplaced. So when we read about the Semantic Web as the next era of the Web, we are less critical of the claims – we do not want to make the same mistake twice. Yet it seems prudent to examine the future of the Semantic Web more

carefully with an eye toward differing perspectives on and expectations of its use, as well as theoretical and pragmatic considerations that will affect its evolution.

The Semantic Web is the outgrowth of many diverse desires and influences, all aimed at making better use of the Web as it stands. The anxiety over the apparent disorder of this new world of digital documents – how one makes sense of new genres, new technologies, and new uses and modes of publishing and organizing materials – is one such influence [24]. A second comes from the field of Artificial Intelligence, with its maturing sense of the kinds of computation that can take place given formal representations – what kinds of problems are tractable to the methods that have been developed over the past 30 years (see for example [27]). Finally, there is a utopian desire to offload the burden of information overload and the complexity of everyday life onto the computer, using the vast resources that have accumulated on the Web as a backdrop to help us in our everyday activities and to address the most normal of problems [5]. All three of these desires and influences are readily justified, given the scope and depth of the information on the Web; we now must ask ourselves which of them are realistic? How can we set appropriate expectations for the reach of the Semantic Web?

From the W3C's inception, there was a perceived need to bring order to the loosely connected networks of digital documents that made up the Web. Although this order was to be realized by consortium's development of standards, it would also reflect the order that libraries have and the Web does not – a consistent structure by which people can access materials. More recently, we can see evidence that this view of the Semantic Web is still widely held in the Hypertext and World-Wide Web communities [8]; Scenario 1 in [29], an information access scenario in which the retrieval is aided by semantic metadata, is a good example.

A second perspective for the Semantic Web is one of a globally distributed knowledge base. This perspective on the Semantic Web was put forth early in the Web's development by Berners-Lee, who began his efforts with the aim of eventually creating networked knowledge ontologies [3]. Berners-Lee has gone on to describe the Semantic Web as being able to learn from the experience of Cyc [23], creating an infrastructure for knowledge acquisition, representation, and utilization across diverse use contexts [4]. In scenarios reminiscent of Apple's Knowledge Navigator vision from the mid 1980's [1], this global knowledge base will be used by personal agents to collect and reason about information, assisting people with tasks common to everyday life.

A third perspective on the Semantic Web is as infrastructure for the coordinated sharing of data and knowledge. In this vision,

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

HT '03, August 26-30, 2003, Nottingham, United Kingdom.

Copyright 2003 ACM 1-58113-704-4/03/0008...\$5.00.

developers create a distributed knowledge or data base for their particular domain-oriented applications. The representation language, the communication protocols, and the access control and authentication are handled by the Semantic Web. This perspective is similar to Bieber and Kacmar's efforts to add computation to hypertext [6], and Halasz's exhortation to this effect in his influential Seven Issues paper [17].

These three perspectives lead to very different expectations of what the Semantic Web will bring to the Web as we know it today. Some of these expectations consider only the technical feasibility and do not consider the social and cognitive implications of the approach, much like Xerox's 1970 vision of a paperless office [36]. Other expectations ignore the difficulty of scaling knowledge-based systems to reason across domains, like Apple's Knowledge Navigator, or are overly optimistic that common sense results from the representation of a sufficient body of domain-oriented knowledge.

The difficulty of knowledge acquisition, representation and reasoning has a long history of being underestimated by some of the field's most influential researchers, including Simon and Minsky:

"Machines will be capable, within twenty years, of doing any work that a man can do." [40]

"... within a generation the problem of creating 'artificial intelligence' will be substantially solved." [30]

This paper analyzes the feasibility of these three general perspectives on the Semantic Web and the expectations that stem from them. In the next section we describe the three perspectives in more detail and provide a framework for examining them. In so doing, we summarize the theoretical challenges of each. We revisit some of our earlier work on bringing formal representations to hypertext, and frame this work in the context of the Semantic Web as a way of anticipating some of its likely challenges. We conclude with a pragmatic look at some of the obstacles the Semantic Web will encounter, discuss two existing Semantic Web applications, and examine some possible near-term outcomes.

## 2. THREE PERSPECTIVES

What is a Semantic Web and what can it do? These are the questions that people may have when they read the articles from the W3C or hear Semantic Web presentations at conferences, meetings, and workshops. High-level visions and scenarios dive quickly into implementation details and standards; it is difficult to sort out what is theoretically and practically possible. To begin the process of sorting out the promise and perils of the Semantic Web, we describe the three perspectives in more detail with examples of how these perspectives are portrayed in writings about the Semantic Web.

Figure 1 places the three perspectives within a space. On one axis, we can think of the representations used on the Web as moving from the particular – limited to the author's original motivation for publishing something on the Web – to the universal, useful in any context. On the other axis, we can consider who uses the representations, human users who are accessing the information directly, either as the result of a query or as the result of interacting with a Web application, or computational processes, which are either knitting together the information holdings of

specific known applications or which are weaving a silent tapestry of knowledge through the work of agents.

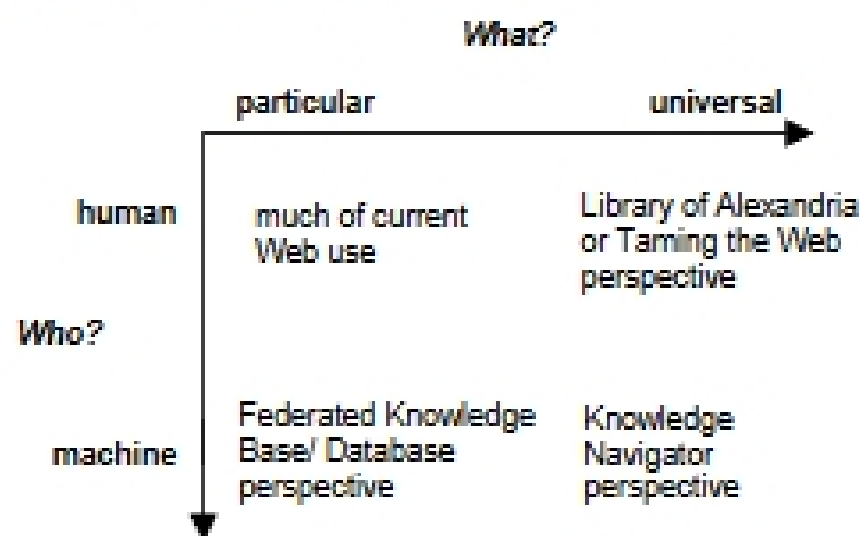


Figure 1. A framework for the three perspectives on the Semantic Web

We can readily put much of current Web use in the realm of the human and the particular. Naturally, the Library of Alexandria or Taming the Web vision that sees the Web's potential to form the ultimate digital library or information resource is further along the axis toward the universal although human use is still the main anticipated outcome. The Federated Data/Knowledge base dwells in the realm of machine processing of the particular and known; through this approach, specific bits of the existing Web are rendered interoperable. In the most distant region of the space lies the Semantic Web writ large, the view that holds it to be the resource of personal agents forming the backdrop for a latter-day Knowledge Navigator.

### 2.1 Taming the Web

One of the early visions of the Semantic Web arose as a reaction to the disorder of the Web. The Web was and is not ordered in a categorization scheme and, until AltaVista and Google came along, seemed to be growing topsy-turvy to the point that the volume of data could not be accessed in an efficient manner. Metadata, cataloging, and schemas were seen as the answer.

With improved indexing and retrieval algorithms, this perspective is rarely discussed any more, although many researchers warn us that search engines are not apolitical [21]. However, human information needs are being met, and the Yahoo's hand-cataloging efforts are in danger of being put out of business by Google's extensive automated index. But remnants of this perspective persist in current writings about the Semantic Web.

"While XML is designed to describe the structure of a document, rather than its content, it is a key tool in two developments aimed at radically improving information retrieval, and in taming the web." [14]

The need for "taming" is no longer the focus of most Semantic Web efforts, although the requirements for current visions make assumptions about the cooperation of authors. Agreeing on a cataloging scheme for Semantic Web documents is a prerequisite for any sharing of semantic knowledge. URIs represent concepts and RDF expresses knowledge as URI triples in the form of (Noun Verb Object). URIs must be used consistently or else the semantics of the concepts will become ill-formed and open for interpretation.

“The Semantic Web is an extension of the current Web in which information is given well-defined meaning, better enabling computers and people to work in cooperation. It is based on the idea of having data on the Web defined and linked such that it can be used for more effective discovery, automation, integration, and reuse across various applications.” [19]

As this quote indicates, the Semantic Web is now viewed as an extension to, rather than a more rigid representation for, the existing Web. This is fortuitous as, while economic reasons might cause businesses to use a prescriptive representation, other sources providing useful information are unlikely to be influenced by such a mechanism.

## 2.2 Knowledge Navigator

In 1987, Apple Computer produced the Knowledge Navigator video in which a personal agent helps a professor deal with incoming messages and his schedule as well as correlating deforestation in South America with the reduction of rainfall in Africa [1]. Aside from the Knowledge Navigator’s natural language interface, this view of a network of knowledge that can be used by personal agents is the primary perspective of many of the current writings about the Semantic Web.

“The Semantic Web will bring structure to the meaningful content of Web pages, creating an environment where software agents roaming from page to page can readily carry out sophisticated tasks for users.” [5]

“Software agents can use this information to search, filter and prepare information in new and exciting ways to assist the web user.” [19]

The view of a machine readable web to go alongside the existing human readable web seems straightforward enough. The difficulty lies in the content of that web. It is easy enough for computers to exchange data about computational abstractions such as filenames, sizes, usernames, passwords, etc. It is much harder for computers to exchange information about human-oriented concepts such as happiness and beauty. These examples are extremes of sorts – but consider the communication that occurs in an on-line book review. There is often a numerical rating that computers can easily share and reason about and there is a written review about why the numerical rating was what it was. Trying to get people to express such a book review in a computer representation will result in limiting what the reviewer can say and making a lot of assumptions about shared understandings.

Furthermore, well-represented concepts in one situation often do not apply to other situations. For example, eBay sellers are given a trust rating; this enables buyers to transfer funds with the confidence that the seller will ship them the item described online, and that it will more or less meet the buyer’s expectations. Trust is accumulated over a number of transactions and seems to work reasonably well in the sprawling eBay virtual garage sale. Now take this quantitative, demonstrated notion of trust, associate it with an identity, and bring it into the Amazon Web site, where readers may contribute their own reviews of books. I may trust a known eBay seller to send me an authentic rare first edition as described on the seller’s page, but would I trust the same eBay seller’s review of the same book on Amazon? Perhaps we should specialize this notion of trust to “material trust” and “intellectual

trust.” We can rely on our eBay seller to deliver the goods, but perhaps not taste or judgment. It is easy to see how general ontologies may spiral out of control, especially as new kinds of goods and services become available (I trust this person’s taste, but not his annotations, for example).

What of data exchanges in the scientific community? Isn’t the exchange of quantitative data and measurements more straightforward? As Star demonstrated in her work with the Worm Community [41], such exchanges prove to be similarly problematic, even without the difficulties that arise when cross-lab sharing is initiated without understanding current work practice.

In general, the intertwined problems of knowledge acquisition, knowledge representation, and knowledge utilization have been the focus of artificial intelligence for close to 50 years. There are now many knowledge-based systems but these are domain-oriented. Their acquisition processes, representations, and use are designed with an understanding of the problems that they will be used to solve. This problem/domain specialization avoids the problems of unending definitions and context and conflicting representations. Minsky [31] and Suchman [42] describe different problems related to representing relevant context. Minsky argues that a single formal representation cannot be used to define what people mean by “being a bird”. His argument proceeds through a variety of definitions of bird with continual counter-examples of where the definition breaks down. In the end, he argues, it is not possible to arrive at such a representation for all circumstances. Suchman describes the problems of a knowledge-based system that lacks sufficient information about the context of use and argues that such gaps are inevitable as there is always a situation where more contextual information is relevant. These are problems with decontextualized knowledge representation. They have also been a continual challenge to attempts to create an encyclopedic knowledge base, e.g. Cyc [23].

In fact, questions about the similarity of the Semantic Web and Cyc have been raised by the W3C. In Berners-Lee’s description of “What the Semantic Web can Represent”, he says:

“... has this not been tried before with projects such as KIF and cyc? The answer is yes, it has, more or less, and such systems have been developed a long way.” [4]

Learning from the experience of Cyc and the Knowledge Interchange Format (KIF) is important for this perspective of the Semantic Web to succeed even partially. The lesson is that context-free knowledge representation relies on domain orientation. Context-aware representations, developed with a specific task in mind, can bring together knowledge that crosses domains. This implies an explosion in the number of URIs for any given concept since different attributes will be important (and have different values) in different contexts. This is not a problem that can be solved by using a uniform knowledge editor, the Semantic Web equivalent to FrontPage. Reasoning in unanticipated ways across the Semantic Web would thus require resolving inconsistencies between different representations and produce highly heuristic results.

## 2.3 Federated Data/Knowledge Base

The last perspective we are considering is that of a federated data/knowledge base. This is similar to the prior perspective except it assumes that the federated components are developed