# 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, 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, 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,


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

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 of 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.

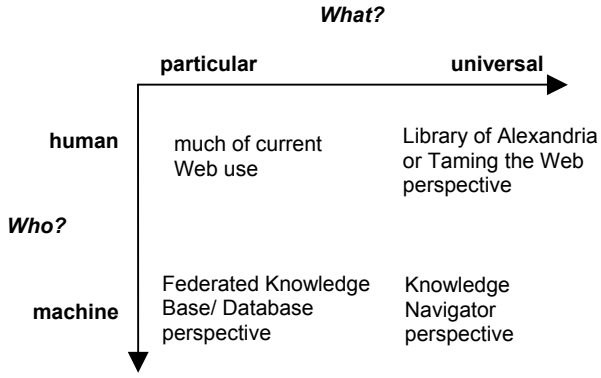


**Figure 1. A framework for the three perspectives on the Semantic Web**

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

with some knowledge of one another or at least with a shared anticipation of the type of applications that will use the data. Much of the existing Semantic Web infrastructure falls into this category – languages used for syntactically sharing data rather than having to write specialized converters for each pair of languages.

> "Indeed, one of the driving forces for the Semantic web, has always been the expression, on the Web, of the vast amount of relational database information in a way that can be processed by machines." [4]

Note that to be successful this perspective requires at least an implicit negotiation about the exchange – what data is represented, and how it is made available by the institutions that are responsible for it.

Information in relational databases may be the primary type of data on the Semantic Web. The following quote from Berners-Lee indicates that he expects simple concepts to be the focus of Semantic Web. This perspective is more likely to succeed than the others since companies will find the reduced costs of maintaining and retrofitting databases to be worth any additional up-front cost.

> "… a large majority of the information we want to express is along the lines of 'a hex-head bolt is a type of machine bolt,' …" [5]

The tendency to extrapolate from the solvable problem of federating existing knowledge bases or databases to viewing every computing-using entity as a potential knowledge source may be somewhat problematic:

> "The Semantic Web will provide an infrastructure that enables not just web pages, but databases, services, programs, sensors, personal devices, and even household appliances to both consume and produce data on the web." [19]

As long as it is clear what information the microwave needs and what useful information it can provide, the interchange representations defined by the W3C are as good or better than others due to the greater likelihood of buy in by a critical mass of companies and software developers. Still, the social processes required for successfully sharing data cannot be avoided [18].

# 3. FORMAL KNOWLEDGE REPRESENTATION CONSIDERED HARMFUL?

Knowledge-based approaches are feasible only if the knowledge can be acquired in a useful format. The Semantic Web provides representation languages that can be used to share that knowledge once it is captured. But when is it possible to acquire formal representations of knowledge and when may these knowledge representations be used? To answer these questions we will first discuss more general, theoretical problems with the use of formal representations and then discuss more pragmatic problems that are found in the context of the Semantic Web.

In our previous article "Formality Considered Harmful" we described the central problems of formal representations to be cognitive overhead, tacit knowledge, premature structure, and situation-specific representations [38]. Much of this discussion was centered on the use of semi-formal representations such as argumentation and schema-based hypertext. In this section we review and update this discussion to better match the questions posed by the context of the Semantic Web.

## 3.1 Additional Overhead

There are a variety of types of overhead that are created by using a more formal representation as opposed to a less formal representation. These include learning the representation and making decisions about how to represent knowledge. This additional effort comes with associated costs in time and expense. Usually the answer to whether this overhead is worthwhile comes from a careful analysis of who bears the overhead – for certainly anything worthwhile is going to entail a certain amount of overhead – and who derives the benefit (see, for example, Grudin's discussion of why groupware applications fail for this reason [16]).

> "… sometimes it is less than evident why one should bother to map an application in RDF. The answer is that we expect this data, while limited and simple within an application, to be combined, later, with data from other applications into a Web." [3]

To encode any knowledge in a formal representation requires the author of that knowledge to learn the representation's syntax and semantics. While learning the basics of HTML is relatively straightforward, learning a knowledge representation language or tool requires the author to learn about the representation's methods of abstraction and their effect on reasoning. For example, understanding the class-instance relationship, or the superclass-subclass relationship, is more than understanding that one concept is a "type of" another concept. The representation implies attributes of one object are also part of another object and that changes to an object's attributes will potentially impact other objects. These abstractions are taught to computer scientists generally and knowledge engineers specifically but do not match the similar natural language meaning of being a "type of" something. Effective use of such a formal representation requires the author to become a skilled knowledge engineer in addition to any other skills required by the domain. Peper and colleagues found that moving from a knowledge-based system to a hypertext reduced costs of maintenance [33]. Good user interface design can mitigate learning the representation's syntax but cannot remove the requirement of learning the representation's semantics.

Once one has learned a formal representation language, it is still often much more effort to express ideas in that representation than in a less formal representation like natural language. One must decide how to subdivide ideas and entities, locate the appropriate concept if it already exists or create it if it does not, name it, and attach it to other concepts, such as attaching attributes or relations to other concepts. Indeed, this is a form of programming based on the declaration of semantic data and requires an understanding of how reasoning algorithms will interpret the authored structures. This effort may be reduced through tools for authoring Semantic Web knowledge representations but these applications will need to be specialized for a context or domain. These systems, much like the high-level programming languages described by Brooks [7], "free a program from much of its accidental complexity" but not the natural complexity of the problem domain.

The added overhead of authoring content for the Semantic Web need not solve the entire formalization problem. Formal representations are already in use for many tasks. Databases

include inventories of parts, accounting software represents accounts payable and receivable, design software represents circuits, VLSI designs, and CAD diagrams. The Semantic Web – as conceived from a federated knowledge/data base perspective – will make it easier to interconnect these already formal representations.

## 3.2 Tacit and Evolving Knowledge

A second problem for working with formal knowledge representations is knowing what to express. People use knowledge that they are not aware of, e.g. tacit knowledge [34]. Requiring this tacit knowledge to become conscious knowledge so it can be represented will, at best, increase the overhead we describe above. But the problem may be worse as it may not be possible for people to express particular concepts.

Formal representations are rigid -- things either are expressed or they are not. This does not match natural communication. This is particularly true with emerging or evolving concepts. When people are working on a problem, their understanding of that problem evolves as they work to solve the problem [11]. Representations created early during such an effort may have to be revised many times during problem solution. Sometimes such knowledge evolution is relatively straightforward, such as the revision of attributes or relations. But often there is a need to revise a whole set of concepts and their interrelations by revising the abstract concept hierarchy. Consider the addition of the microwave oven to the class of ovens in the 1970s. A microwave oven serves the same purpose (to heat things), but uses such different methods that the concept hierarchy for ovens will most likely have to change, resulting in the creation of additional abstract classes to express these similarities and differences [10]. End-user modifiability methods support non-knowledge engineers in activities such as the placement of new objects in an existing ontology [15]. These techniques have met with limited success in practice, perhaps because they too can only reduce accidental complexity for their user.

The W3C understands that the evolution of knowledge representations is a requirement although they are more concerned with issues of knowledge transfer and maintenance that are discussed in more detail in the next section.

> "A requirement of namespaces work for evolvability is that one must, with knowledge of common RDF at some level, be able to follow rules for converting a document in one RDF schema into another one (which presumably one has an inate [sic] understanding of how to process)." [3]

The problem of tacit and evolving knowledge will impact the expression of individual users more than the description of commercial content. For this reason the Semantic Web will find it easier to support access to content about products and services than more abstract materials.

## 3.3 Situated Nature of Knowledge

Knowing what to express and what to leave unsaid is one of the great arts of common sense. Rather than enumerating all the facts about an object or topic, we decide which are important to our current context. Similarly, knowledge representations also must decide on bounds of activity. Experience indicates that the more focused the domain or problem of application, the better the knowledge-based approach is likely to work.

Rosch's study of natural categorization shows that people give different names (and attributes) to the same objects in different contexts [35]. At home or the office we may call what we sit on a "chair". In a furniture store we might call it an "office chair". In an office supply store, we might call it by its brand name (Aeron) or mention attributes of the chair (ergonomic). While this is easily represented as a hierarchy of concepts ranging from general ("chair") to more specific (model number and attributes), the characteristics we emphasize will depend on who we are talking to and what we are talking about.

Knowledge representation is similarly problem dependent. In the examples from Minsky and Suchman that we referred to earlier, the attributes for an object and the context are not enumerable. Merging representations developed by people in similar or different contexts will require a lot of understanding and reasoning about the contexts of use. Consistency may require many representations for same concept and lack any way of knowing for which contexts they are similar enough or how to convert between them without human intervention.

> "Where for example a library of congress schema talks of an 'author', and a British Library talks of a 'creator', a small bit of RDF would be able to say that for any person x and any resource y, if x is the (LoC) author of y, then x is the (BL) creator of y." [3]

This is an example of such a conversion. In this case, we assume that the British Library definition of "creator" is a superset of the Library of Congress definition of "author". After all, painters, sculptors, and even programmers are creators. But even this definition can get murky. When a well-known photographer takes a picture of a famous building, is the photographer or the architect the creator? Creation is an abstract concept that will require agreement among parties if consistency is to be maintained. Within a limited context, such as libraries containing written materials, such agreement may be possible.

# 4. PRAGMATIC ISSUES

Naturally the Semantic Web raises pragmatic issues as well as theoretical ones. To identify and explore these issues, we can approach them from three different angles: (1) the Semantic Web's viability in the growing arena of metadata initiatives, standards, and practices; (2) the Semantic Web's reliance on the informed adoption of emerging mark-up standards; and (3) the Semantic Web's strategic efficacy and utility, given competitors such as Google and its ilk, which may work well enough to transform the web into a universal information resource, and Amazon and its kin, which may be sufficiently effective in roles as commercial middlemen and transaction mediators.

Each of these three angles will shed light on different aspects of the Semantic Web – how it will be built and who will bear the cost, whether existing mark-up practices support assumptions about the Semantic Web's widespread adoption, and what the Semantic Web's de facto competitors are likely to be. We see these three perspectives as pragmatic since we can use the considerable body of experience with today's Web as a crystal ball to foresee the shape of the emergence and trajectory of the Semantic Web.

In each of these discussions, we take as our foil the universal perspective on the Semantic Web rather than the particular. It is interesting to note that many of these pragmatic issues may be

addressed by taking an intelligent systems perspective that limits and focuses the approach on particular resources and specific requirements for interoperability.

## 4.1 Looking at the Semantic Web as Metadata

From the perspective of a librarian, cataloger, publisher, or content provider, the Semantic Web is a metadata initiative; at the heart of the Semantic Web is the assumption that adding formal metadata that describes a Web resource's content and the meaning of its links is going to substantially change the nature of the way computers and people find material and use it. Because there are a variety of metadata efforts underway – that is, the Semantic Web is a metadata initiative among many – it is important to evaluate the Semantic Web in this context. There are three aspects of Web metadata to consider: (1) community; (2) cost; and (3) authority or trust. Although these three aspects are intertwined, they may be pulled apart for the purpose of discussion.

### *4.1.1 Metadata and Community*

Metadata is not simply a description of the information contained in a work or web page; the choice of a metadata scheme also signifies community membership. Every aspect of metadata – from how it is obtained and verified to the expectations of how it will be used by humans or computer systems – stems from the practices of a particular community. What metadata is necessary? Who creates it? Does it ever come from the people who use the information resource, possibly guided by a tool? Or must it arise from a professional cataloger? May it be created by automatic analysis? What are the sources of authority for metadata values? Is it sufficiently uniform to warrant federation across different organizations? It is easy to see that these are questions that rely on community membership.

For example, the adoption of online library catalogs has resulted in a standardized metadata schema, MARC (Machine Readable Cataloging format). As is the case with other metadata formats, MARC promotes cross-institutional standardization and search interoperability. It reflects and propagates the cataloging practices of a community. Not only that, but the community's coding practices for setting metadata values are well documented; in the United States, catalogers typically adhere to the rules specified by the AACR (the Anglo-American Cataloging Rules) and agreed upon methods of authority control. Records for commonly held books and information resources may be obtained from an external concern (OCLC) and – by using standards like Z39.50 – searches may cross organizational boundaries and different online catalogs (OPACs). Dublin Core, an effort to standardize metadata for Web resources is a similar kind of initiative, but it is slow to make visible progress, due in part to some of the arguments we have put forth here.

It is also important to consider the underpinnings of a community's metadata coding practices. Again, using our library example, good librarianship demands minimal interpretation of the work to arrive at the appropriate metadata values, so the cataloging is consistent across institutions. Of course, some amount of interpretation is unavoidable, and catalogers will readily admit that no set of cataloging rules is ever fully prescriptive. There is always an occasion for human judgment. But it currently appears that the Semantic Web will rely extensively on human interpretation and judgment to bring metadata values into conformance with the ontology, and in fact, to derive and extend the ontology in the first place.

### *4.1.2 Metadata Cost*

To complicate matters further, each of the many competing metadata standards and initiatives has an associated cost. Will any community actually have the financial wherewithal to conform to a jumble of multiple standards? For example, the National Science Digital Library efforts in the United States use metadata as a means of federating distributed resources; metadata from individual collections is automatically harvested and brought together in a centralized resource that uses OAI (Open Archives Initiative), an XML-based metadata schema [22]. Will the collection developers – some of them instructors with little extra time and resources to devote to cataloging – code more than one kind of metadata? At the current time this looks doubtful. Some of the most valuable and authoritative resources on the Web will need to adhere to community standards and will not be able to bear the cost of generality.

One argument for the Semantic Web is that the cost will be borne by many; each collection developer and content creator will play his or her part in tagging the Semantic Web. Yet professional catalogers have long noted that this is a problematic approach to reducing cost [26]. Users tend to catalog information resources with an implicit (and often inconsistent) sense of how they themselves intend to use the resource. They are neither trained to apply authority control, nor do they have a larger sense of the institutional purpose of a given resource.

In some cases, the cost of using multiple mark-up or metadata standards has been reduced by using translation scripts and rules. For example, the University of Virginia's Electronic Text Center prepares an initial version of electronic texts and e-books using TEI (Text Encoding Initiative) tags, since TEI is a community standard for humanities scholars. They then convert this version to other popular formats (e.g. HTML or Open E-book) through an evolving set of scripts [13]. Will this strategy work for the Semantic Web? Seemingly not, since the Semantic Web is in many ways orthogonal to other metadata standards that mainly address functional parts of a Web document. Furthermore, tagging links with semantics has proven to be notoriously difficult in the past (see for example, the evolutionary path from NoteCards to Aquanet to VIKI [25]); while we might think that semantic links are possible and desirable, they are certain to be costly to code and verify.

### *4.1.3 Metadata Authority and Trust*

Earlier in this section, we claimed that a sense of community gives metadata initiatives context (how and when the metadata will be used) and sets expectations for the trustworthiness of the metadata (its degree of accuracy and consistency, and whether a searcher feels confident that the results reflect a desired balance of precision and recall). One set of perspectives on the Semantic Web is that it is universal, in which case it would not rely on a similar sense of community. Rather, it would extend across communities, and gain its power from its ubiquity and its ability to transcend particular applications and specialized uses. From this perspective, the Semantic Web's strong emphasis on representational expressiveness, extensibility, and flexibility places no limitations on who might be creating Semantic Web metadata and why. Establishing trust – that the metadata is a good

and consistent representation of content for the use to which it is put – will be a challenge.

On the other hand, if we move toward the less general Federated Knowledge/Data Base perspective on the Semantic Web which emphasizes its strength in bridging between particular resources with a well-defined purpose in mind, authority and trust become less of an issue. In this case, visible evidence of reliable performance over time may ameliorate any trust issues associated with intelligent systems behavior, but they are certain to be there in early phases of use.

## 4.2 Looking at the Semantic Web as Mark-up

The Semantic Web is a mark-up based solution for adding semantically meaningful metadata to the Web; at its core, we find XML and RDF, both embedded mark-up representations of a Web document's content. Thus looking at HTML in action may help us predict the outcome of semantic mark-up.

Was HTML a success? At first glance, this seems like an absurd question. Of course HTML was a success – every Web page uses it; it has created the demand for special-purpose editors such as Macromedia's Dreamweaver or Microsoft's FrontPage; and most common text editors such as Microsoft Word will produce it as a format for publishing documents. How can it be considered anything less than a success?

Indeed, from the standpoint of ubiquity of adoption and familiarity, HTML has been an unparalleled success. It is simple and robust; many writers learned how to mark up documents in HTML to good end. But it is also important to understand why it worked. When HTML was introduced, the expectation was that it would be used as a simple language for expressing the functional structure of a document, much more approachable and readily applied than its sophisticated predecessors like SGML. Thus it emphasized simple document elements such as titles, headers, paragraphs, and unordered lists, and did not require authors to define the Document Type Definitions (DTDs) that would correspond to the many document genres its designers expected to see on the Web. Document appearance was originally left to the browser. In short, HTML was the great egalitarian mark-up language.

What happened? Although the very early Web documents followed the spirit of HTML, authors quickly began to care what their documents looked like to readers and to make the Web page genre immediately apparent to readers [12][37]. They wanted margins, so they went through all manner of trickery to indent text an inch from the edge of the browser window (using, for example, the infamous blank images), and eventually settled on more complicated ways of achieving this effect as HTML began to include additional functional elements like tables. In short, inspection showed that few Web pages actually used HTML as functional mark-up. <H2> tags appeared without <H1>s for example. Eventually, Web designers resorted to manipulating the fonts directly: why specify an unpredictable <H1> if all you really want is a text string that's Times Roman 14 point bold? Designers worked to get the appearance they wanted, rather than trying to represent the invisible abstraction of a document's functional structure.

Did Web design software make the situation better? Indeed not. The HTML that these editors produced often ends up with little functional meaning. The simplicity of the original HTML design is hidden by complicated paragraph-by-paragraph markup. A "view source" on professionally designed commercial Web pages reveals a vertiginous tangle of tags aimed solely at producing the desired visual effect and the placement of advertising. A "View Source" on any authoritative newspaper site (see for example www.nytimes.com) confirms this emphasis on good visual layout and consistent appearance. Although the adoption of XML and CSS stylesheets may have some effect on this practice, it is doubtful that the new standards will be adopted in a manner that goes beyond or conflicts with the look and "branding" of this kind of resource.

The question now is, will the Semantic Web and the editors designed for producing Semantic Web mark-up transcend this problem? In some cases, we expect that the mark-up will reflect a profound commitment to the Semantic Web and its principles. But in others, we must expect the analogous result to occur: Semantic Web mark-up will be good enough to solve the immediate problem or produce the desired behavior in a limited range of high payoff situations.

## 4.3 Looking at the Semantic Web as its own "Killer App"

According to its proponents, the Semantic Web will fundamentally change the way we interact with the Web; "the Semantic Web is the killer app" [5]. Many of the activities in the scenarios used to illustrate the power of the Semantic Web are accomplished today with some degree of success using Google, individual Web applications or services, or by bypassing the computer altogether. Let's examine one of the scenarios used in the Scientific American article (see Figure 2).

> The entertainment system was belting out the Beatles' "We Can Work It Out" when the phone rang. When Pete answered, his phone turned the sound down by sending a message to all the other *local* devices that had a *volume control*. His sister, Lucy, was on the line from the doctor's office: "Mom needs to see a specialist and then has to have a series of physical therapy sessions. Biweekly or something. I'm going to have my agent set up the appointments." Pete immediately agreed to share the chauffeuring.
>
> At the doctor's office, Lucy instructed her Semantic Web agent through her handheld Web browser. The agent promptly retrieved information about Mom's *prescribed treatment* from the doctor's agent, looked up several lists of *providers*, and checked for the ones *in-plan* for Mom's insurance within a *20-mile radius* of her *home* and with a *rating* of *excellent* or *very good* on trusted rating services. It then began trying to find a match between available *appointment times* (supplied by the agents of individual providers through their Web sites) and Pete's and Lucy's busy schedules. (The emphasized keywords indicate terms whose semantics, or meaning, were defined for the agent through the Semantic Web.)
>
> In a few minutes the agent presented them with a plan. Pete didn't like it—University Hospital was all the way across town from Mom's place, and he'd be driving back in the middle of rush hour. He set his own agent to redo the search with stricter preferences about *location* and *time*. Lucy's agent, having *complete trust* in Pete's agent in the context of the present task, automatically assisted by supplying access certificates and shortcuts to the data it had already sorted through.
>
> Almost instantly the new plan was presented: a much closer clinic and earlier times—but there were two warning notes. First, Pete would have to reschedule a couple of his *less important* appointments. He checked what they were—not a problem. The other was something about the insurance company's list failing to

include this provider under *physical therapists*: "Service type and insurance plan status securely verified by other means," the agent reassured him. "(Details?)"

**Figure 2. A Semantic Web scenario quoted from [5]**

In the scenario, Lucy's agent looks for a specialist while she is at the doctor's office arranging an unspecified treatment for her mother. Why wouldn't Lucy simply ask the doctor for a referral, then call the recommended specialist to find out whether the specialist is on Mom's insurance plan? It's possible that such a conversation might even change the insurance status of the specialist – the insurance situation in the United States is so complicated that many doctors now rely on their patients to guide them into joining a particular preferred provider organization. But let's suspend disbelief momentarily and find an alternative non-mediated way to find the specialist. First, it's likely that they'll be doing additional research on Mom's condition anyway rather than relying on whatever information the doctor has assembled, since it is now common for people – both clinicians and patients – to look up medical information online [20]. It is also likely that they'll simply use the online version of Mom's provider directory to decide which specialist to use; note that the search requires exactly the same information that Pete and Lucy are giving to the agent. A particular solution of the federated database/knowledge base type is more likely to work here, one that puts together patient-understandable knowledge of Mom's condition with map locations and specialist names.

The idea that Pete and Lucy will actually represent all the nuances of their busy schedules is unlikely; even they themselves are not apt to be able (or willing) to articulate the relative importance of various events scheduled on their calendars [32]. Sharing calendars outside of one's immediate social circle (i.e. having patients or patients' agents schedule their own appointments) is even more problematic [32]. It is also unlikely that Pete will keep his Web-controlled agent in sync with his acoustic desires; if he's of a mind to have his answering machine pick up the call, he's unlikely to want his stereo turned down. It is difficult to keep detailed preferences in line with moment-to-moment needs.

In general, what we've seen on the Web is a tension between the ability of content and service providers to create high-quality metadata that anticipates user needs and the ability of users themselves to represent their needs to construct good queries that do not rely on extra metadata. This tension can be caricatured as the difference between Yahoo's approach, which relies on human cataloging skills, and Google's approach, which relies on clever automated processing and the user's ability to articulate his or her information needs within a helpful interface.

The Semantic Web is closer to Yahoo's approach. In situations in which user needs are known and distributed information resources are well described, this approach can be highly effective; in situations that are not foreseen and that bring together an unanticipated array of information resources, the Google approach is more robust. Furthermore, the Semantic Web relies on inference chains that are more brittle; a missing element of the chain results in a failure to perform the desired action, while the human can supply missing pieces in a more Google-like approach. Although the information retrieval research literature readily acknowledges that it is difficult for people to construct good queries, it easier for them to converge on an acceptable answer through reformulation – even naïve reformulation – than it is to get the knowledge representation that the Semantic Web relies on to work right in all situations. Google takes advantage of the intrinsic properties of Web pages (including social evaluation through linking, file format, topical categorization, and URL structure) rather than relying on the availability of extrinsic metadata and complicated constructed ontologies; in many cases, this strategy addresses many of the same issues associated with using the Web as a universal library.

Do pragmatic considerations always work against the Semantic Web? No, but scenarios of the complexity of Figure 2's Knowledge Navigator-like approach to interacting with people and things in the world seem unlikely. Similarly, the "taming the Web" approach violates what we know about the costs and problems with metadata creation. On the other hand, cost-benefit tradeoffs can work in favor of specially-created Semantic Web metadata directed at weaving together sensible well-structured domain-specific information resources; close attention to user/customer needs will drive these federations if they are to be successful.

# 5. DISCUSSION OF EXISTING SEMANTIC WEB APPLICATIONS

In Section 4.3, we analyze some pragmatic difficulties implicit in a scenario written from the futuristic Knowledge Navigator perspective. Current work on the Semantic Web also includes work more typical of the Taming the Web perspective [39] and applications that are good illustrations of the Federated Knowledge Base/Database perspective [2]. To explore how the issues we introduced earlier in this paper manifest themselves in current applications, we will discuss briefly two examples: semantic relationships between scholarly documents, and mappings between classification schemes in e-commerce.

The first application, Buckingham-Shum and colleague's ScholOnto [39], uses Semantic Web representations to enhance the existing scholarly practice of citing related work by explicitly identifying relationships of the paper to other papers and physical or conceptual entities. The example from this application shows a hypertext researcher specifying these relationships for the Dexter Hypertext Reference Model. The benefit of such specification is in aiding later researchers to uncover what motivated this work and its impact.

The expression of inter-document/object relations within the hypertext community raises a number of the issues previously identified. First, who enters this information? The motivation for expressing these relationships is unclear even in the case of academics, a community picked by the authors for being more likely to express such relations. Second, how is the content entered deemed trustworthy by other members of the community? If a person searching the knowledge space must analyze the validity of each relationship, then automatic citation analysis by tools like CiteSeer, which show the textual context of a citation, provide similar results without the knowledge acquisition cost.

The second example illustrates a method for solving a canonical B2B problem, integrating and reconciling suppliers' catalogs whose contents have been self-described using two different current e-commerce standards, the American standard (UNSCSP) and the European standard (ECL@SS). The solution involves a semi-automated method for performing the mapping using a tool,

SI-Designer [2]. This example serves as a good illustration of the kind of application the Semantic Web is designed to address in the near term, federating existing, in-use databases. These databases are likely to be accessible today through individual Web applications.

Certainly, with sufficient human intervention, this solution is apt to work and addresses many of our pragmatic issues of metadata community, cost, and authority. However there is no magic, and certainly the theoretical issues associated with federating enormous product catalogs and mapping among different e-commerce standards is going to be a laborious process on the part of the human counterparts of the electronic middlemen even given some amount of automation. For example, in the authors' scenario, it is not clear whether thesauri and other knowledge sources employed in the mapping will need to be extended to contain the common sense knowledge about the differences between American letter-sized paper and European A4 paper. Nor do abstract consumer expectations of toilet paper translate well across international boundaries. The devil here is indeed in the situated details, and these details add up.

What these examples show is that even in what appear to be specialized applications of semantic web representations there is substantial room for the same types of problems found in the more domain-independent scenario described in Figure 2.

# 6. CONCLUSIONS

What should we expect from the Semantic Web? What is within the grasp of the emerging standards and methods, given the breadth of experience of Computer and Information Science research? Perhaps the first perspective, that of taming the web to create a universal digital library – a modern-day Library of Alexandria – is more in the realm of Google-like approaches that do not rely on such an abrupt shift in the practices and economies surrounding the Web. The second perspective, the one that finds the Semantic Web in the role of a true Knowledge Navigator, seems out of reach for both theoretic and pragmatic reasons.

What we are left with then is the Semantic Web's potential in the realm of the particular, especially its ability to weave together specific resources well motivated by business (in the now-famous B2B or B2C arenas) or institutionally-supported digital libraries (for example, in medicine, where both clinicians and patients are accessing and using very complicated resources to make informed decisions [20]). In these cases, resources may be available to perform the kinds of knowledge engineering, cataloging, and librarianship that is called for. Domain-specific ontologies and agents operating within prescribed kinds of human activities, using specialized mark-up (for example, DAML-S [28]), are good candidates for a Semantic Web approach.

Whenever we look toward the Semantic Web and its promise, we must remember to consider the very basic theoretical and practical questions:

- Knowledge stability (How well are the domain and the practices surrounding it understood? How much incremental formalization and restructuring do we expect?)
- Competing conceptual approaches (Is the knowledge intrinsic or extrinsic? Can intrinsic structure be recognized through heuristic approaches, thus avoiding declared representations?)
- Cost/benefit (Who will do the knowledge representation, and to what end? What are competing interests, e.g. other metadata standards?)
- Negotiation among information resource stakeholders (What is the role of negotiation, facilitation, or intervention in representing the knowledge within a socio-technical framework? Are there identified and accepted approaches that work in the domain, e.g. [9]?)

Answering these basic questions will allow us to adopt Semantic Web standards, methods, and practices, and bring its computational power to an appropriate set of human activities. It may be that – at least in the short term – that there are many semantic webs rather than The Semantic Web; they may – even in the long term – take us where we need to go.

# 7. ACKNOWLEDGEMENT

This work was supported in part by National Science Foundation grants No. 9734167 and 0226321. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation or Microsoft Corporation.

# 8. REFERENCES

[1] Apple Computer (1992, May). Knowledge navigator. In B. A. Myers (Ed.), *Conference on Human Factors in Computing Systems: Special Video Program*, ACM/SIGCHI.

[2] Bergamaschi, S., Guerra, F., and Vincini, M. A Data Integration Framework for e-Commerce Product Classification. *Proc. First International Semantic Web Conference* (Sardinia, Italy, June 9-12, 2002), 379-393.

[3] Berners-Lee, T. Semantic Web Road map. http://www.w3.org/DesignIssues/Semantic.html (last modified Oct. 14, 1998).

[4] Berners-Lee, T. What the Semantic Web can represent. http://www.w3.org/DesignIssues/RDFnot.html (last modified Sept. 17, 1998).

[5] Berners-Lee, T., Hendler, J., and Lassila, O. The Semantic Web. *Scientific American*, May 17 2001, 34-43.

[6] Bieber, M. and Kacmar, C. Designing hypertext support for computational applications. *Communications of the ACM* August 1995, 38 (8), 99-107.

[7] Brooks Jr., F.P. No Silver Bullet: Essence and Accidents of Software Engineering. *IEEE Computer*, 20(4), (April 1987), 10-19.

[8] Carr, L., Hall W., Bechhofer S., Goble C. Conceptual linking: ontology-based open hypermedia, *Proceedings of the tenth international conference on World Wide Web*, 2001, 334-342.

[9] Conklin, J., Selvin, A., Shum, S.B., Sierhuis, M. Facilitated hypertext for collective sensemaking: 15 years on from gIBIS. In *Proceedings of Hypertext 2001*, New York: ACM Press, 2001, 123-124.

[10] Fischer, G., and Girgensohn, A. End-User Modifiability in Design Environments. In *Human Factors in Computing Systems, CHI.90 Conference Proceedings* (Seattle, Wash., April). ACM, New York, 1990, 183-191.

[11] Fischer, G., and Reeves, B.N. Beyond Intelligent Interfaces: Exploring, Analyzing and Creating Success Models of Cooperative Problem Solving. *Applied Intelligence, Special Issue Intelligent Interfaces 1*, (1992), 311-332.

[12] Furuta, R., Marshall, C.C. "Genre as Reflection of Technology in the World-Wide Web" in *Proceedings of the International Workshop on Hypermedia Design* (June 1-3) Montpellier, France, LIRMM 1995, 203-214.

[13] Gibson, M. and Ruotolo, C. Beyond the Web: TEI and the Ebook Revolution. *Proceedings of ACH'01*. http://www.nyu.edu/its/humanities/ach_allc2001/papers/gibson/index.html.

[14] Gilchrist, A. From Aristotle to the 'semantic web', *Record of the Library Association*, January 2002, http://www.la-hq.org.uk/directory/record/r200201/article2.html

[15] Girgensohn, A. End-User Modifiability in Knowledge-Based Design Environments. Ph.D. Diss.., Department of Computer Science, University of Colorado, Boulder, Colorado, 1992.

[16] Grudin J. Why CSCW applications fail: problems in the design and evaluation of organization of organizational interfaces, *Proceedings of the 1988 ACM conference on Computer-supported cooperative work*, 85-93.

[17] Halasz, F.G. Reflections on NoteCards: Seven Issues for the Next Generation of Hypermedia Systems. *Communications of the ACM*, July 1988, 31(7), 836-852.

[18] Hatala, M. and Richards, G. Global vs. Community Metadata Standards: Empowering Users for Knowledge Exchange. *Proceedings of the First International Semantic Web Conference* (Sardinia, Italy, June 9-12, 2002), 292-306.

[19] Hendler, J., Berners-Lee, T., and Miller, E. Integrating Applications on the Semantic Web, 2002, http://www.w3.org/2002/07/swint.

[20] Hersh, W. Medical Informatics: Improving Health Care Through Information. *Journal of the AMA*, October 23/30, 2002, Vol. 288, No. 16. 1955-1958.

[21] Introna, L. and Nissenbaum, H. Shaping the Web: Why the Politics of Search Engines Matters. *The Information Society*, 16(3), 2000, 1-17.

[22] Lagoze C. and Van de Sompel, H. The open archives initiative: building a low-barrier interoperability framework. In *Proceedings of ACM/IEEE-CS Joint conference on Digital libraries*, 54-62.

[23] Lenat, D. B. "Cyc: A Large-Scale Investment in Knowledge Infrastructure." *Communications of the ACM* 38, no. 11 (November 1995).

[24] Levy, D.M. *Scrolling Forward: Making Sense of Documents in the Digital Age*. New York: Arcade Publishing, 2001.

[25] Marshall, C.C., Shipman, F.M. III. "Spatial Hypertext: Designing for Change." *Communications of the ACM*, 38, 8 (August 1995), 88-97.

[26] Marshall, C. "Making Metadata: a study of metadata creation for a mixed physical-digital collection" in *Proceedings of the ACM Digital Libraries '98 Conference*, ACM, 162-171.

[27] McDermott, D. and Dou, D. Representing Disjunction and Quantifiers in RDF. *Proceedings of the First International Semantic Web Conference* (June 9-12, 2002), 250-263.

[28] McIlraith, S. and Martin, D. L. Bringing Semantics to Web Services. *IEEE Intelligent Systems*, 18, 1, 90-93.

[29] Miles-Board, T., Kampa, S., Carr, L., Hall, W. Capturing Meaning: Hypertext in the semantic web. In *Proceedings of Hypertext 2001*, ACM, 2001, 237-238.

[30] Minsky, M. *Computation: Finite and Infinite Machines*, Prentice-Hall, 1967, p. 2.

[31] Minsky, M. A Framework for Representing Knowledge. In *The Psychology of Computer Vision*, P. Winston, Ed. McGraw-Hill, New York, 1975, 211-277.

[32] Palen, L. Social, individual and technological issues for groupware calendar systems. In *Proceedings of CHI'99*, ACM Press: New York, 17 - 24. 1999.

[33] Peper, G., MacIntyre, C., and Keenan, J. Hypertext: A New Approach for Implementing an Expert System. *Proceedings of ITL Expert Systems Conference*, 1989.

[34] Polanyi, M. *The Tacit Dimension*. Garden City, NY: Doubleday, 1966.

[35] Rosch, E. Principles of categorization. in: Rosch, E.; Lloyd, B. B. (eds.), *Cognition and Categorization*. New Jersey: Hillsdale, 1978.

[36] Sellen, A. and Harper, R. *The Myth of the Paperless Office*. Cambridge: MIT Press, 2001.

[37] Shepherd, M. and Watters, C. The Functionality Attribute of Cybergenres. in *Proceedings of HICCS'99-Volume 2*, January 5-8, 1999, Maui, Hawaii, 2007-2015.

[38] Shipman, F. and Marshall, C. Formality Considered Harmful: Experiences, Emerging Themes, and Directions on the Use of Formal Representations in Interactive Systems, *Computer Supported Cooperative Work (CSCW)*, 8, 4 (Fall 1999), 333-352.

[39] Shum, S.B., Motta, E. and Domingue, J. ScholOnto: An Ontology-Based Digital Library Server for Research Documents and Discourse. *International Journal on Digital Libraries* (2000), 3 (3), 237-248.

[40] Simon, H. *The Shape of Automation for Men and Management*, New York: Harper & Row, 1965.

[41] Star, S.L. and Ruhleder, K. 1996. "Steps towards an ecology of infrastructure: Design and access for large-scale collaborative systems," *Information Systems Research*, 7, 111-138.

[42] Suchman, L. *Plans and Situated Actions: The Problem of Human-Machine Communication*. Cambridge, UK: Cambridge University Press, 1987.