INDEX

SYNOPSIS

1. Semantics is arguably the single most important ingredient in propelling the Web to its next phase, and is closely supported by Web services and Web processes that provide standards based interoperability of applications. Semantics is considered to be the best framework to deal with the heterogeneity, massive scale, and dynamic nature of the resources on the Web.

2. The semantic web is an approach to facilitate communication by making the web suitable for machine-to-machine communication. It can be use to encode meaning and complex relationships in web pages. A major challenge for the emerging semantic-web field is to capture the knowledge required and structure it a format that can be processed automatically (eg by agents).

3. Appropriate tools can assist humans in populating the semantic web with annotations and create models for intermediate adapter systems that allow intelligent agents to make use of ontologies and other information on the semantic web.

4. Currently, the web is a gigantic, mainly static, source of information. This situation means that the heavy burden in information access, extraction, and interpretation is left to the human web users. Means to put machine-understandable data on the web are becoming increasingly important. The web has been said to reach its full potential only when it becomes a place where data can be shared, processed, and understood by automated systems as well as by people.

INTRODUCTION

"So far as the laws of mathematics refer to reality, they are not certain. And so far as they are certain, they do not refer to reality."

- Albert Einstein

1. Technology is supposed to make life simpler .That includes making it unnecessary for us to rack our brains when we're conducting a search on the web. There should be no need to be an 'expert Googler', if there were such a term. However, most of today's search engines actually do require that - that you frame your query well. If you're searching for a local place that serves up pizza, and you'd like to order online, you should be able to just type in 'pizza' and get your restaurant. That would entail some level of localization capabilities built into the search engine, and also the fact that the engine should be able to determine what your intent is—namely, that you want to buy a pizza.

2. If much of the problem with finding the right page lies in the information that the pages give out to search engine crawlers, the answer might just lie in the vision of Tim Berners-Lee called the Semantic Web, in the context of the World Wide Web (WWW).

3. From Semanticweb.org, "the Semantic Web is a vision: the idea of having data on the web defined and linked in a way that it can be used by machines." It is, essentially, a project that aims to create a more intelligent Web by annotating pages on the Web with their semantics (meaning), in a manner understandable by computers (or search engine spiders). Thus, if a spider knew what a page was about, it would return more relevant results.

4. Using XML and other technologies, this information can be made explicit in the page that contains these elements. As of now, the Semantic Web is only a vision, with its proponents and detractors.

AIM

5. To describe the emergence of the "Semantic Web".

PREVIEW

6. For better understanding and easy assimilation, the technical paper would be covered in the following four parts: -

(a) Part I : What is Semantic Web.
(b) Part II : Evolution and Architecture.
(c) Part III : Semantic Web Technologies.
(d) Part IV : Applications.

.

PART - I : WHAT IS SEMANTIC WEB

7. 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. (Tim Berners-Lee, James Hendler and Ora Lassila).

8. The Semantic Web is a vision: the idea of having data on the web defined and linked in a way that it can be used by machines not just for display purposes, but for automation, integration and reuse of data across various applications. (http://www.w3.org/2001/sw/).

9. The Semantic Web is an initiative of World Wide Web consortium (W3C).

(a) Semantic Web is a set of Languages and Tools for machine processing of information stored in WWW.

(b) It is an efficient way of representing data on the World Wide Web, or as a globally linked Knowledge Base.

(c) Semantic Web is about an efficient Knowledge Representation (KR) mechanism for Artificial Intelligence (AI).

(d) Semantic Web is about efficient Reasoning Systems required for integration of distributed data.

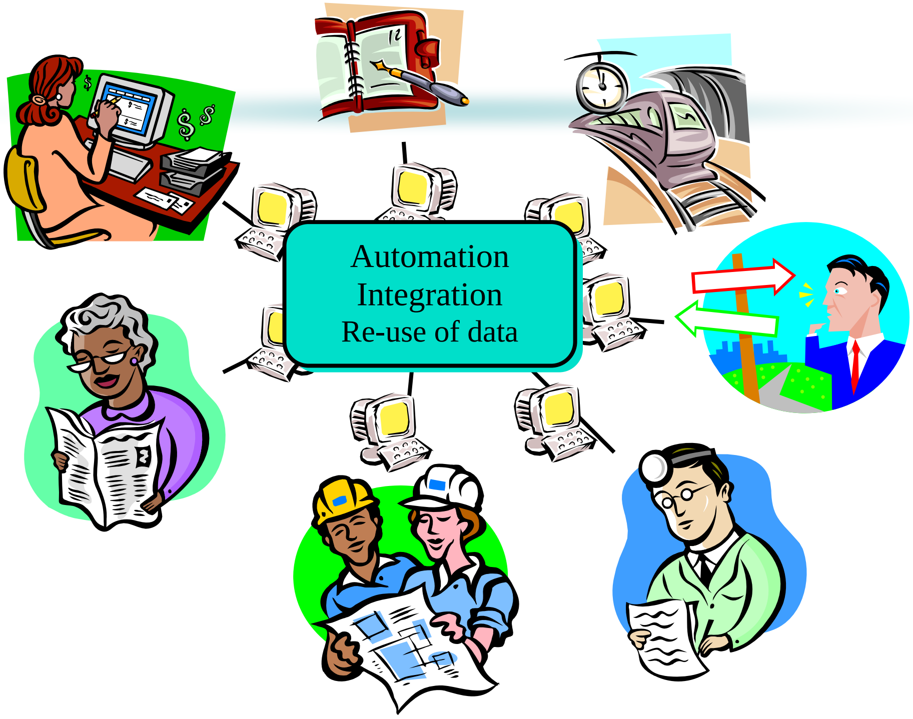


Fig 1: Semantic Web

10. The Semantic Web is a mesh of information linked up in such a way as to be easily processable by machines, on a global scale. It is an efficient way of representing data on the World Wide Web, or as a globally linked database (fig 1).

11. The Semantic Web was thought up by Tim Berners-Lee, inventor of the WWW, URIs, HTTP, and HTML. There is a dedicated team of people at the World Wide Web consortium (W3C) working to improve, extend and standardize the system, and many languages, publications, tools and so on have already been developed. However, Semantic Web technologies are still very much in their infancies, and although the future of the project in general appears to be bright, there seems to be little consensus about the likely direction and characteristics of the early Semantic Web.

12. On the WWW, data that is generally hidden away in HTML files is often useful in some contexts, but not in others. The problem with the majority of data on the Web that is in this form at the moment is that it is difficult to use on a large scale, because there is no global system for publishing data in such a way as it can be easily processed by anyone. For example, consider the information about local sports events, weather information, plane times and television guides, all of this information is presented by numerous sites, but all in HTML. The problem with that is that, is some contexts, it is difficult to use this data in the ways that one might want to do so.

13. So the Semantic Web can be seen as a huge engineering solution or even more than that. As it becomes easier to publish data in a repurposable form, more people will want to publish data, and there will be a knock-on or domino effect. Consequently a large number of Semantic Web applications can be used for a variety of different tasks, increasing the modularity of applications on the Web.

14. The Semantic Web is generally built on syntaxes which use Uniform Resource Identifiers (URIs) to represent data, usually in triples based structures: i.e. many triples of URI data that can be held in databases, or interchanged on the World Wide Web using a set of particular syntaxes developed especially for the task. These syntaxes are called Resource Description Framework (RDF) syntaxes.

PART – II : EVOLUTION AND ARCHITECTURE

15. Let us first see the Semantic Web in the context of the evolution of the Internet. In the beginning, Internet resources were stored on FTP (File Transfer Protocol) servers. If it was known which FTP servers existed, one could browse through the directory structures and download files whose names resembled the topics being searched for. To provide a location mechanism, Gopher was invented. Gopher servers organized resources by category. Humans classified the resources. A search facility based on descriptive information was provided. Each Gopher described the resources on its own site, commonly belonging to a university or some other institution. A significant breakthrough was the Veronica server, a Gopher server that knew of the existence of most of the world's gopher servers and could route the user to the Gopher servers that were most likely to describe the information that was being sought.

16. The World Wide Web superceded this. When the WWW was small, the directory at CERN could describe all the Web Sites in the world and provide a hyperlink for each. As the web exploded, new and more sophisticated directory systems emerged. Yahoo is an example of a well-known directory system. Yahoo is a database that classifies information on the Web within a hierarchy of topics. This classification is human driven - it relies on the skill of a small army of editors, experts in their domain, each responsible for a certain topic. A search in Yahoo searches the descriptive data, not the contents of the data itself.

17. The hyperlink mechanism of the WWW facilitated the growth of search engines. Search engines consists of two parts. The first part is the crawler that reads documents indexes, their text contents and follows links to other documents, bringing more and more documents into their knowledge bases. The second part is the search system that accepts a user request and uses the text index to retrieve documents that contain word patterns suggesting that the document may be relevant.

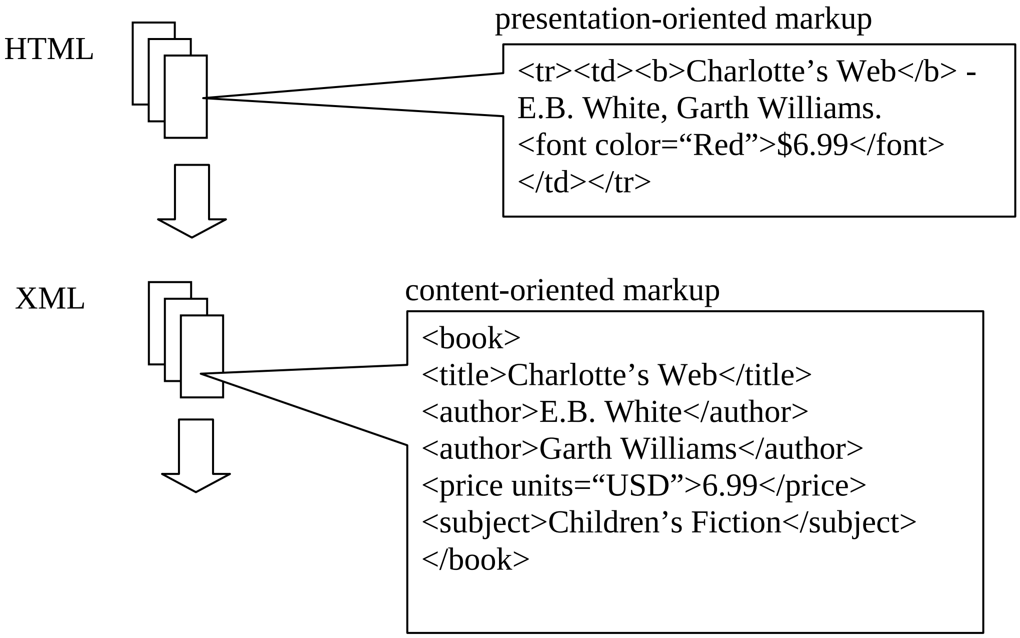
18. There are some hybrid systems. These search engines attempt to classify documents by topic and meta search machines (for example the Meta Crawler). The latter are engines that will submit a user request to many search engines and aggregate the results. The technical limitations of these systems include the following :-

(a) There is no uniform interface to directory systems or search engines. They are designed for use by humans.

(b) The metadata in directories is minimal. At best, there is a classification by topic, a title, a description, and a hyperlink to the resource.

(c) Search engine indexes contain the text indices for the document which are resources but the existence of the word "Tolstoy" does not indicate if the document was about Tolstoy, written by Tolstoy, published by Tolstoy, was reviewing a book by Tolstoy, advertising a book by Tolstoy or providing an application allowing purchasing a book by Tolstoy. It is not possible to tell if the "Tolstoy" in the page is the Tolstoy who wrote Anna Karenina or some other person or animal with the same name. If a human retrieves the page, he or she might be able to work this out. However, a computer cannot.

19. Over time web standards evolved from Hyper Text Markup Language Standards (HTML) which were presentation oriented to Extensible Markup Language Standards (XML) which are content oriented (fig 2).



```
<tr><td><b>Charlotte’s Web</b> -
E.B. White, Garth Williams.
<font color=“Red”>$6.99</font>
</td></tr>
```

```
<book>
<title>Charlotte’s Web</title>
<author>E.B. White</author>
<author>Garth Williams</author>
<price units=“USD”>6.99</price>
<subject>Children’s Fiction</subject>
</book>
```

Fig 2: Evolution of Web Standards

20. The Semantic Web timeline is as illustrated in fig 3. These are a series of initiatives by World Wide Web Consortium (W3C).

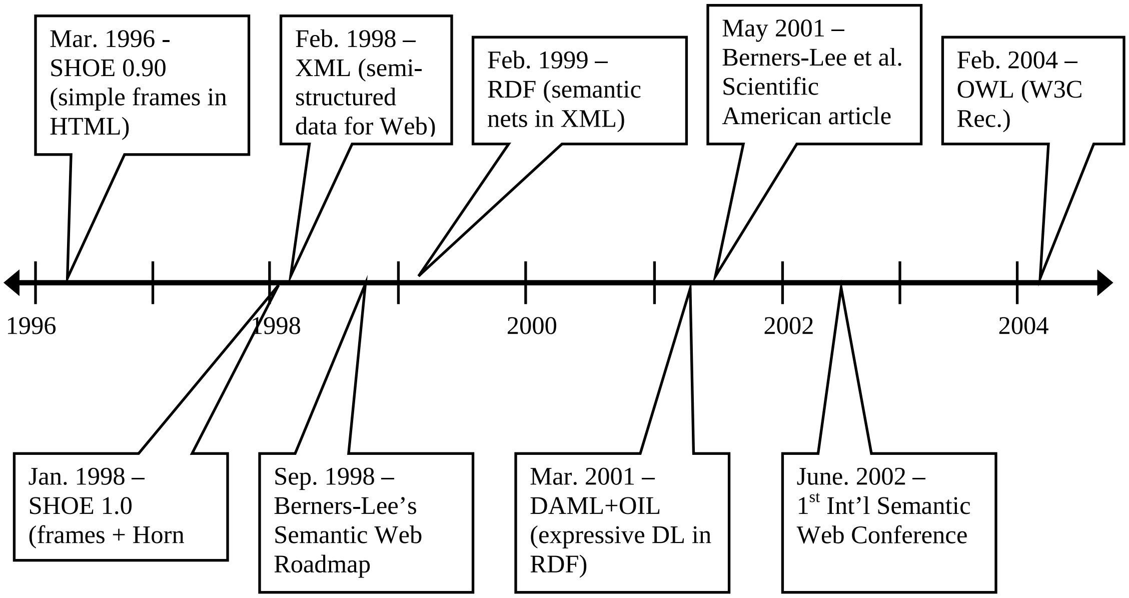

Mar. 1996 - SHOE 0.90 (simple frames in HTML)
Feb. 1998 – XML (semi-structured data for Web)
Feb. 1999 – RDF (semantic nets in XML)
May 2001 – Berners-Lee et al. Scientific American article
Feb. 2004 – OWL (W3C Rec.)
1996
1998
2000
2002
2004
Jan. 1998 – SHOE 1.0 (frames + Horn
Sep. 1998 – Berners-Lee's Semantic Web Roadmap
Mar. 2001 – DAML+OIL (expressive DL in RDF)
June. 2002 – 1st Int'l Semantic Web Conference

Fig 3: Semantic Web Timeline

21. Fig 4 illustrates a box architecture of the Semantic Web. The terms used have been elaborated in the succeeding paragraphs.

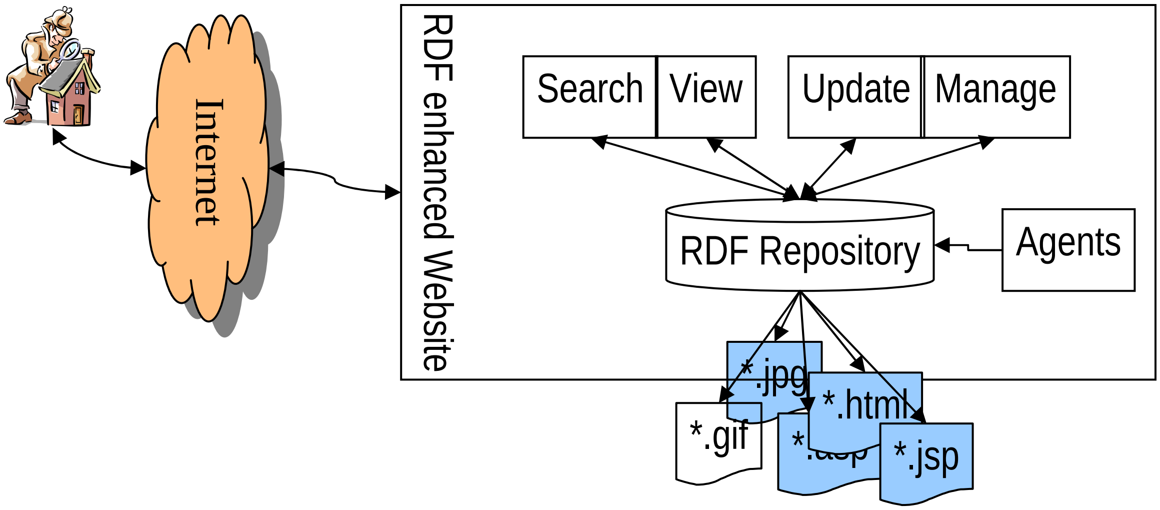


Fig 4: Semantic Web In a Box Architecture

22. Fig 5 illustrates the various layers of the Semantic Web as conceived by its pioneer Tim Berners-Lee of the W3C. Up to the ontology layer the layers are in the implementation phase and upwards from there are in the research phase. Most of these terms are discussed in depth in this paper.

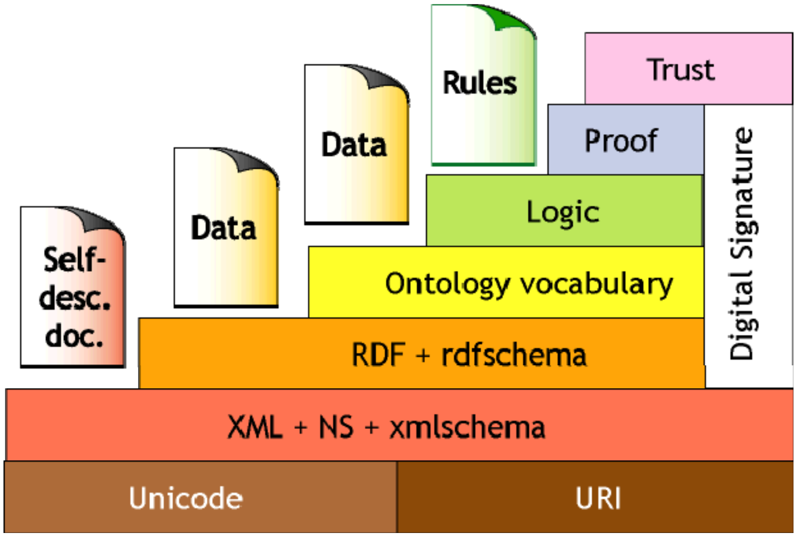


Fig 5: Semantic Web Architecture

23. A schema and an ontology are ways to describe the meaning and relationships of terms. This description (in RDF) helps computer systems use terms more easily, and decide how to convert between them. The W3C is preparing to start a Web Ontology (WebOnt) Working Group (which some call the WOW-G). This group is chartered to prepare a Web Ontology language that builds upon the work done by RDF Schema and DAML+OIL.

24. Logic. Using RDF based systems basic concepts like subclass, inverse, etc can be understood but it would be even better if any logical principle could be stated and the computer could reason (by inference) using these principles. Consider this example. One company decides that if someone sells more than 100 of their products, then they are a member of the Super Salesman club. A smart program can now follow this rule to make a simple deduction: "John has sold 102 things, therefore John is a member of the Super Salesman club."

25. Proof. Once systems that follow logic are built, it makes sense to use them to prove things. People all around the world could write logic statements. Then any machine could follow these Semantic "links" to construct proofs. Consider this example. Corporate sales records show that Jane has sold 55 chains and 66 sprockets. The inventory system states that chains and sprockets are both different company products. The built-in math rules state that 55 + 66 = 121 and that 121 is more than 100. And

someone who sells more than 100 products is a member of the Super Salesman club. The computer puts all these logical rules together into a proof that Jane is a Super Salesman. It is however, very difficult to create these proofs (it can require following thousands, or perhaps millions of the links in the Semantic Web), but it's very easy to check them. In this way a Web of information processors is built. Some of them merely provide data for others to use. Others are smarter, and can use this data to build rules. The smartest are "heuristic engines" which follow all these rules and statements to draw conclusions, and kindly place their results back on the Web as proofs, as well as plain old data.

26. Trust and Digital Signature. Based on work in mathematics and cryptography, digital signatures provide proof that a certain person wrote (or agrees with) a document or statement. So a person digitally signs all of his RDF statements. This way, a computer is sure about who wrote the documents (or at least can vouch for their authenticity). Now, a program is simply told whose signatures to trust and whose not to. Each can set their own levels or trust .The computer can decide how much of what it reads, to believe. A computer may be told to trust a friend, say Robert. Robert happens to be a rather popular guy on the Net, and trusts quite a number of people. And of course, all the people he trusts, trust another set of people. Each of those people trust another set of people, and so on. As these trust relationships fan out from you, they form a "Web of Trust." And each of these relationships has a degree of trust (or distrust) associated with it.

27. Note that distrust can be as useful as trust. Suppose the computer discovers a document that no one explicitly trusts, but that no one explicitly distrusts either. Most likely, the computer will trust this document more than it trusts one that has been explicitly labeled as untrustworthy. The computer takes all these factors into account when deciding how trustworthy a piece of information is. It can also make this process as transparent or opaque as is desired. For example, someone might be happy with a simple "thumbs up/thumbs down" display. Someone else might insist on a complex explanation, including a description of some or all of the trust factors involved in the decision..

PART – III : SEMANTIC WEB TECHNOLOGIES

EXTENSIBLE MARKUP LANGUAGE (XML) AND RESOURCE DESCRIPTION FRAMEWORK (RDF)

28. One of the fundamental contributions towards the Semantic Web to date has been the development of XML . Liberating data from opaque, inextensible formats as it does, XML provides an interoperable syntactical foundation upon which solutions to the larger issues of representing relationships and meaning can be built. It is an important center of agreement among individual developers and corporations. The face of the Web is changing, offering once again new possibilities for communication and interaction - not because all of the underlying concepts are new *per se*, but because they can be combined on the Web and exposed to the opportunity and unpredictability of large-scale decentralization.

29. By now, XML is widely known in the WWW community and is the basis for a rapidly growing number of software development activities. XML is intended as a markup language for arbitrary document *structure*, as opposed to HTML, which is a markup language for a specific kind of hypertext documents. An XML document consists of a properly nested set of open and close tags, where each tag can have a number of attribute-value pairs. Crucial to XML is that the vocabulary of the tags and their allowed combinations is not fixed, but can be defined per application of XML. An example serializes a part of the ontology given above.

```
<class-def>

<class name="plant"/>

<subclass-of>

<NOT><class name="animal"/></NOT>

</subclass-of>
```

```
</class-def>
<class-def>
<class name="tree"/>
<subclass-of>
<class name="plant"/>
 </subclass-of>
</class-def>
<class-def>
<class name="branch"/>
<slot-constraint>
<slot name="is-part-of"/>
<has-value>
<class name="tree"/>
</has-value>
</slot-constraint>
</class def>
```

30. From the indentation of the above example, it is easy to see that the basic data-model of XML is a labeled tree, where each tag corresponds to a labeled node in the data-model, and each nested sub-tag is a child in the tree. It is important to point out that the above is only one possible XML syntax for the ontology given above. Many other XML versions of the same semantic information are also possible. The possibility for

multiple serializations stems from the fact, that XML is foremost a means to define grammars. Different grammars can be used to define the same syntactic content.

31. Perhaps surprisingly, a powerful tool for the construction of the Semantic Web is HTML itself or, more properly, XHTML. Most people are acquainted with the "meta" tags which can be used to embed metadata about the document as a whole. Yet there are more powerful, granular techniques available too. Although largely unused by web authors, XHTML offers several facilities for introducing semantic hints into markup to allow machines to infer more about the web page content than just the text. These tools include the "class" attribute, used most often with CSS stylesheets. A strict application of these can allow data to be extracted by a machine from a document intended for human consumption. For instance, consider the example:

```
<p>
  For more information, contact:
  <span class="contact" id="edumbill">
    <span class="name">Edd Dumbill</span>,
    <span class="role">Managing Editor</span>,
    <span class="organization">XML.com</span>
  </span>
</p>
```

32. A program could easily construct from such a XHTML snippet a "Contact" object identified by the ID "edumbill" with properties "name", "role" and "organization”. Techniques similar to this, known colloquially as "screen scraping," have been used for some time on the Web. Common applications include the extraction of data from search engines for use in Perl scripts or the extraction of headline information from news sources. For these applications the problem has been the shifting nature of the design of HTML pages and, thus, the need to readjust the scrapers whenever the design changes. A page marked up using the technique showed above would enable reliable scripts to interface with the HTML.

33. As web application providers consider adding Simple Object Access Protocol (SOAP) and similar interfaces to their systems to allow remote-application access, they

could actually be saved the effort of maintaining twin APIs (browser and SOAP) by embedding machine-readable information in the HTML itself. There is still a lot of value and utility in simpler web technologies.

34. Once the richer information has been embedded in a page, a program still needs to transform it into the format it requires. At this point another W3C technology, XSLT, has a lot to offer. Given an XHTML page as input, it is useful for selecting and transforming the contents of that page. It provides an excellent bridge from older HTML technology to the nascent XML-based Semantic Web applications. A tool of singular utility when used in conjunction with an XSLT processor is "Tidy," which can take HTML and turn it into XHTML. As most web authoring tools still don't have XHTML support, HTML will be created by web authors for some time to come. Tidy facilitates the processing of normal HTML with XSLT, enabling authors of such documents to participate in the Semantic Web.

35. The Semantic Web Activity also provides the Resource Description Framework (RDF). The essence is to provide a convention for the association of properties to resources. Their properties are not part of the resource but a part of a description of this resource.

36. The value of any property can be either a *literal*, like a piece of text or a resource reference (a Uniform Resource Identifier or URI). The name of a property has an identifier that can be unique on the Internet; that is to say the name of this property stands for a meaning which is the same everywhere.

37. Suppose a document is given the property "title" and the literal value "W3C Semantic Web Activity". This does not define what title means. In an English dictionary, there are many meanings of the word "title". It could mean the title of a book, a person’s title (Dr., Mr., Mrs. etc), a degree of ennoblement (Duke, Earl, Count) and so on. How to agree on what "title" represents on a global basis?

38. The answer is to take the property name from a *namespace* and the namespace corresponds to the words in a *vocabulary*. So users or machines that share the vocabulary will interpret the meaning of a property based on the RDF information. RDF as a concept was invented before XML appeared on the scene but XML was rapidly adopted to express the serialized form of RDF (that is, how RDF is written in a document or exchanged between machines).

39. Here is a piece of RDF in XML format that assigns properties to a document :-

```
<?xml version="1.0" ?>
<RDF xmlns="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:dc="http://purl.org/dc/elements/1.1/">
<Description about="http://www.w3.org/2001/sw/">
<dc:title>W3C Semantic Web Activity</dc:title>
<dc:description>This document describes the Semantic Web Activity at the
World Wide Web Consortium.</dc:description>
<dc:publisher>World Wide Web Consortium</dc:publisher>
</Description>
</RDF>
```

xmlns="http://www.w3.org/1999/02/22-rdf-syntax-ns#" declares the default namespace of the XML to be the RDF namespace. The about attribute points to the resource - http://www.w3.org/2001/sw/. It is the subject of the description. The namespace of the document properties is defined by xmlns:dc="http://purl.org/dc/elements/1.1/".

40. This says that any property with a name *somename* written like as “dc: *somename*” belongs to the namespace whose identifier is http://purl.org/dc/elements/1.1/. This namespace belongs to the Dublin Core Metadata Initiative [2], which over several years, has built up a vocabulary to describe documents. This was long before XML and RDF were thought of. These include definitions for title, description, publisher, and several other useful properties. An example of one such definition, designed to be read by a human is as follows:

Element: Title

Name: Title

Identifier: Title

Definition: A name given to the resource.

Comment: Typically, a Title will be a name by which the resource is formally known.

41. In effect what the RDF information says can be enunciated by three statements.

(a) The Dublin Core title property of the resource http://www.w3.org/2001/sw is the literal value "W3C Semantic Web Activity”.

(b) The Dublin Core description property of the resource http://www.w3.org/2001/sw is the literal value "This document describes the Semantic Web Activity at the World Wide Web Consortium".

(c) The Dublin Core publisher property of the resource http://www.w3.org/2001/sw is the literal value "World Wide Web Consortium".

42. ALL RDF information is fundamentally formed as subject, predicate, and object triples(fig 6). This formulation is familiar to logicians and programmers who deal with logical relations For instance, in Prolog the first statement may be written as:

title("http://www.w3.org/2001/sw","W3C Semantic Web Activity**")**

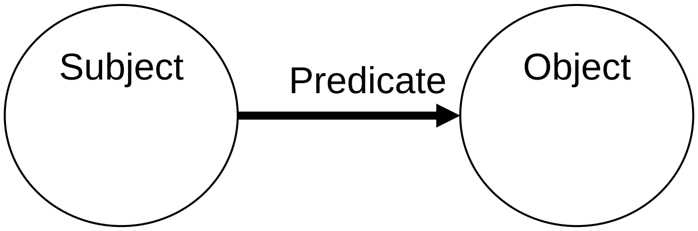


Fig 6: RDF Triple

43. The RDF XML shown here is merely an XML representation of these statements. Alternative XML representations of these statements are also possible. For instance,

dc:title could be an attribute of the Description element instead of a child element. RDF does not care about the distinction between elements and attributes.

44. RDF Interfaces are supposed to be able to process all valid XML RDF representations and to convert them into some internal representation that is extracted from the user, for instance in a database or a prolog program.

45. RDF property values can also be URIs . This permits the expression of relations between resources. As an example,

(a) The document at URI1 is a corrected version of the document at URI2**.**

(b) The document at URI1 is a **c**ommentary on the document at URI2.

(c) The document at URI1 can also be found at URI2. By making statements about statements, one can express more complex relations.

46. The RDF specification is fairly complex. There is an RDF Schema specification which describes how to describe RDF vocabularies in a machine-readable way, rather than the human readable way quoted above for Dublin Core. Unsurprisingly this is also modeled in RDF. The target of the Semantic Web is a global meta-information system and it is based on RDF.

47. RDF is in fact much more useful than this. It corresponds to a very common design pattern for XML usage in real systems. Its relative insensitivity to whether properties are expressed as XML elements or attributes and the sequence of property declarations make it very flexible. “Stand-off Markup” leaves no footprints on the document it describes. This fact makes it usable in a large number of scenarios where the document schema cannot be altered or is not in XML.

48. RDF has already met some criteria which HTML has already possesses. This indicates that RDF is able to scale up to a worldwide phenomenon. These criteria are listed

(a) It is fairly simple to implement, despite a rigorous specification.

(b) If access to the RDF Schema data is not available, the mechanisms that process RDF information still work.

(c) The technology is useful at a fairly local or trivial level.

(d) There are many people working on it.

(e) It fulfills a need.

(f) It has no central point of failure

(g) It is based on fundamental principles, which are in fact not new.

(h) It is composable. RDF systems can refer to other RDF systems using the normal URI resolution mechanism.

(j) RDF processing mechanisms are free to work in the way the implementers think best because there is a layer of abstraction between the RDF Interface and the implementation.

49. Representing Knowledge on the Web with XML or RDF. The Web is a universal medium for exchanging data and knowledge: for the first time in history we have a widely exploited many-to-many medium for data interchange. This medium poses new requirements for any format used for exchanging data on the web:

(a) Universal expressive power. Since it is not possible to anticipate all potential uses, a data format must have enough expressive power to express any form of data.

(b) Support for Syntactic Interoperability. Syntactic interoperability means how easy it is to read the data and get a representation that can be exploited by applications. For example, software components like parsers or query APIs should be as reusable as possible among different applications. Syntactic interoperability is high when the parsers and APIs needed to manipulate the data are readily available.

(c) Support for Semantic Interoperability. Semantic interoperability means the difficulty of understanding the data. Please note the difference from syntactic interoperability: syntactic interoperability talks about parsing the data, while semantic interoperability means to define mappings between unknown terms and known terms in the data. This requirement as one of the most important, since the cost of establishing semantic interoperability is usually higher than that for establishing syntactic interoperability, due to the need for content analysis.

ONTOLOGIES

50. Ontology is a set of knowledge terms, including the vocabulary, the semantic interconnections and some simple rules of inference and logic, for some particular topic. For example the ontology of cooking and cookbooks includes ingredients, how to stir and combine them, the difference between simmering and deep-frying, the expectation that the products will be eaten or drunk, that oil is for cooking in or consuming and not for lubrication, and so forth. More complex logics and inference systems are generally considered as separate from the ontology per se.

51. Recently, a number of research groups have been developing languages in which to express ontological expressions on the web . In an effort to bring these together, and to try to arrive at a *de facto* web standard, a number of researchers, supported by the US Defense Advanced Research Projects Agency (DARPA) released a draft language known as the DARPA Agent Markup Language (DAML). Since then, an ad hoc group of researchers has formed the "Joint US/EU committee on Agent Markup Languages" and released a new version of the language called DAML+OIL. OWL (Ontology Web Language) is also an offshoot of this product which is gaining increasing popularity(fig 7). This language is based on the Resource Description Framework (RDF) and discussion of its features is conducted on an open mailing list (archived at http://lists.w3.org/Archives/Public/www-rdf-logic/). DARPA is now funding a set of researchers both to develop freely available tools, and to provide significant content for these tools to manipulate – thus demonstrating to the government and other parts of society that the semantic web can be a reality, not just a vision.

52. A crucial aspect of creating the semantic web is to make it possible for a number of different users to create the machine-readable web content without being logic experts.

In fact, ideally most of the users shouldn't even need to know that web semantics exists. Lowering the cost of markup isn't enough – for many users it needs to be free. That is, semantic markup should be a by-product of normal computer use. Much like current web content, a small number of tool creators and web ontology designers will need to know the details, but most users will not even know ontologies exist.

53. As an example, consider any of the well-known products for creating on-line slide shows. Several of these products contain libraries of clippings that can be inserted into the presentations. These clippings can be marked with pointers to ontologies by the developers of the software, and the saving of the products on the web (*Save as HTML…*), could include the linking of these saved products to the ontologies they're marked from. Thus a presentation that had pictures of, for example, a cow and a donkey, would be linked to barnyard animals, mammals, animals, etc. While this would not guarantee appropriate semantics (the cow might be the mascot of some school or the donkey the icon of some political party), retrieval engines would be able to use the markups as clues to what are in the presentations and how they may be linked to other ones. The user simply creates a slide show as is done today, but the search tools do a better job of finding it based on content.

54. An alternative example would be a markup tool driven from one or more ontologies. For example, consider a homepage-creation tool that is driven by representing hierarchical class relations as menus. Properties of the classes could be tied to various types of forms, and these made available via simple web forms. A user could thus choose from a menu to add information about a person, and thence choose a relative (vs. friend, professional acquaintance, etc.) and thence a daughter. The system could then use the semantics to retrieve the properties of daughters specified in the ontology/ies and to display them to the user as a form to be filled out by filling in strings (like name), numbers (age); or to browse for related links (homepage), on-line images (photo-of), etc. The system would then lay these out using appropriate web tools, while also recording the relevant instance information.

55. Since the tool could be driven from any ontology, a library of terms could be created (and mixed) in any number of different ways. Thus, a single easy-to-use tool would allow the creation of home pages (using ontologies on people, hobbies, etc.), professional pages (using ontologies relating to specific occupations or industries), agency-specific pages (using ontologies relative to specific functions) etc. In an easy, interactive way a user would be assisted in creating a page and get the markup created for free. Notice, also, that mixtures of the various ontologies and forms could be easily created, thus helping to create the semantic web of pages linking to many different ontologies, as mentioned earlier.

56. Not only can pages be created with links to numerous ontologies, but also the ontologies themselves can include links between them to reuse (or change) terms. The notion of creating large ontologies by the combination of components is not unique to the semantic web vision . However, the ability to link and browse ontological relations enabled by the web's use of semantics will be a powerful tool for those users who do know what ontologies are and why they should be used.

57. Formal ontologies in description logic based representation; supported by deductive inference mechanisms are the primary (but certainly not the only) means of addressing major challenges in realizing the Semantic Web vision.

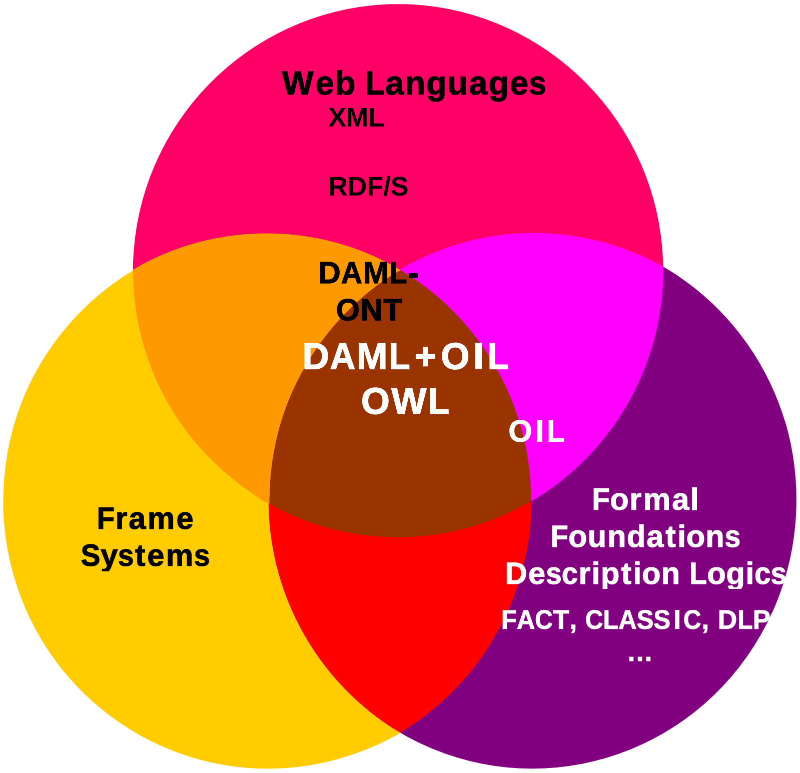


Fig 7: Ontologies

AGENTS ON THE SEMANTIC WEB

58. One of the most powerful uses of the web ontologies described here, and a key enabler for agents on the web, is in the area of web services. Recently, numerous small businesses, particularly those in the area of supply-chain-management for B2B e-commerce, have been discussing the role of ontologies in managing the machine-to-machine interactions. In most cases, however, these approaches have assumed that the ontologies are primarily used by the constructors of computer programs to make sure that they agree on terms, types, constraints, etc. Thus, the agreement is recorded

primarily off line, and used in web management applications. The semantic web goes much further than this, creating machine-readable ontologies used by "capable" agents to find these web services and automate their use.

59. Using a combination of web pointers, web markup, and ontology languages, a can situation much better than just putting service advertisements into ontologies arises. Rather, these techniques will also include a machine-readable description of a service (as to how it runs) and some explicit logic describing the consequences of using the service. In fact, it is these latter two properties – a service description and a service logic – that would lead u to the integration of agents and ontologies in some exciting ways.

60. Consider intelligent web agents, using the analogy of travel agents who would find possible ways to meet user needs, and offer the user choices for their achievement. Much as a travel agent might gives a list of several flights , or a choice of flying versus taking a train, a web agent should offer a slate of possible ways to get you what is needed on the web.

61. Consider the following example (a web-enabled method for saving the doomed crew of *The Perfect Storm*). In this story, now a major motion picture, a crew of fishermen is out at sea when weather conditions conspire to create a storm of epic proportions. For various reasons, the crew is unable to get a detailed weather map, and thus miss the fact that the storm is developing right in their way. Instead of avoiding it, they end up at its center with tragic results. How could web agents have helped?

62. As the Captain of the ship goes to call land, a wave hits and his cell-phone is swept overboard. Luckily, he is a savvy web user, and has brought his wireless web device with him as well. Checking the weather forecast from a standard weather site, he determines that a storm is coming, but he does not find enough detail for his planning needs. He goes to an agent-enabled geographical server site and invokes the query "Get me a satellite photo of this region of the Atlantic (and draws a box on an appropriate map)." The system comes back a little later with the following message displayed.

Query Processed**:**

- A satellite image taken yesterday at 10 AM is available on the web at http://…
- A new satellite image, to be taken today at 10AM, will be available for $100 – click here to authorize transfer of funds and obtain image (you will need a valid credit card number from one of the following providers: …)
- In an emergency situation, a Coast Guard observer plane can be sent to any location within the area you indicate. Service Note: You will be responsible for cost of flight if the situation does not result in emergency pickup. Click here for more information.
- A high altitude observer can be sent to your location in 13 hours. Click here to initiate procedure. (You will need to provide military authorization, a valid military unit code, and the name of commanding officer. Abuse of this procedure can result in fine or imprisonment).
- A service entitled commercial service for providing satellite images is advertised as becoming available in 2004. See http://… for more information.

Results of processing agent-based query

Options range from a picture available on the web (possibly out of date) to other services (that may need special resources) and even future options currently being announced. The captain now chooses an option depending on what resources he has available and what criterion he is willing to accept. Recognizing the gravity of his situation, he invokes the Coast Guard option, and an overflight is scheduled for his GPS location. Seeing the emerging weather, the Coast Guard arranges an emergency pickup at sea, and the sailors are able to go on to fish again some other day. Using the tools of the semantic web, this sort of thing can be made routine and available to any user who needs to use a web service for any purpose by enabling expressive service capability advertisements to be made available to, and usable by, agents on the web(fig 8).

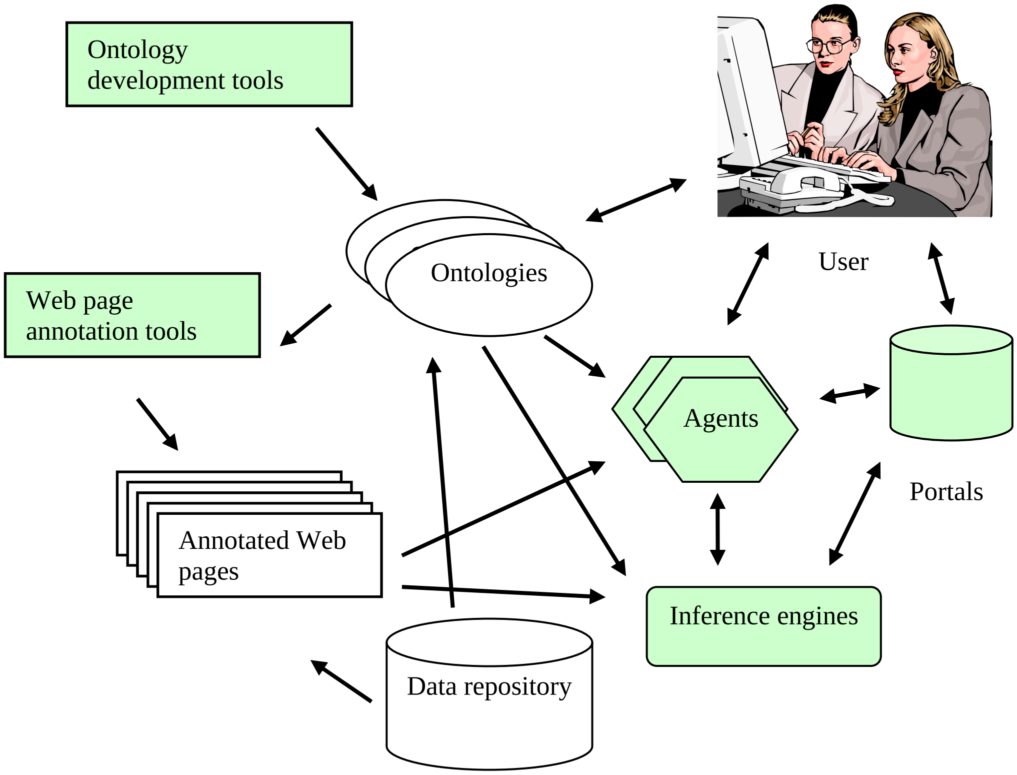


Fig 8: Smart Search

LATENT SEMANTIC INDEXING

63. Another important technique for ‘sematicising’ the web is Latent Semantic Indexing (LSI). With a regular keyword search, a document, looked at from the search engine’s point of view either contains a keyword or doesn’t. There’s no middle ground. And each document stands alone—there’s no interdependence between documents. The Web is not viewed for the collection of documents that it is: it is viewed as a lot of

individual documents taken separately. In LSI, the regular recording of what words a document contains is done first. The important addition is that it examines the document collection as a whole to look for other documents that may contain the same words in a certain document. What does this do? Essentially, if two documents have a lot of words in common, they are ‘semantically close’. And if two documents don’t have many words in common, they are semantically distant.

64. Now a search is performed on a database that has been indexed by the LSI method, the search engine looks for documents that semantically match the keywords. For example, in the semantic system, we’re talking about, ‘crocodile’ and ‘alligator’ are pretty close, so a search on ‘crocodile’ would also bring up pages that contain only ‘alligator’ with no mention of ‘crocodile’.

65. If search engines were to use LSI, they would be more powerful. A search engine looks for pages that contain all your keywords. If twenty keywords were entered into a regular search engine, you’d get very few results—but LSI shows the reverse behaviour. If you enter more search terms into a search engine that has LSI’d the Web, it’s likely to find more, not less, documents of relevance—for the simple reason that it would bring up closely related documents for each keyword. Filtering the results according to relevance, would provide feedback to the engine about what you think best matches your query. This, combined with Personalization, would lead to your results getting much better over time.

ARTIFICIAL INTELLIGENCE AND NATURAL LANGUAGE PROCESSING

66. AI could help determine one’s intent when one feeds in keywords, and it could help in understanding the contents of a page as well. When both these happen, what you have is a smarter search engine. This could happen through implementations of NLP (Natural Language Processing). It’s basically a method by which a computer processes something said in a ‘natural language’ such as English, as opposed to a computer

language, and comes up with something intelligent. That is, when you apply NLP to something like “the world is round”, the machine would have an internal representation of that fact. That sentence would not remain four, un-understood words, but would mean something to the machine. They would mean that something called ‘the world’ has a property, and that that property is called ‘round’. The system could also, if the knowledge has been fed in, know what round means; it might be able to deduce that the world is therefore in some way similar to a ball, and so on. NLP is one of the hardest AI problems. However, Tom Mitchell, former president of the AAAI (American Association of Artificial Intelligence), said in November 2003 that in three to five years, we could have something like this.

67. An example of a current search engine that has AI claims to fame is Accoona. From the “Artificial Intelligence” link on Accoona: “Accoona Artificial Intelligence is a Search Technology that understands the meaning of search queries beyond the conventional method of matching keywords. This user-friendly technology, merging online and offline information, delivers more relevant results and enhances the user experience. Accoona’s AI uses the meaning of words to get you better searches. For example, when you type five keywords in a traditional search engine, you’re going to get every page that has all five keywords, no more, no less. With Accoona’s AI Software, which understands the meaning of the query, the user will get many additional results. Accoona’s AI also supertargets your search. For example, within a query of five keywords, Accoona AI allows the user to highlight one keyword, and will rank the search results starting by every page where the meaning of that one keyword is more important than the meaning of the other four keywords.”

PART -IV: APPLICATION

68. Semantic search and contextual browsing. Consider an ontology consisting of general interest areas with several major categories (News, Sports, Business, Entertainment, etc.) and over 16 subcategories (Baseball, Basketball, etc in Sports). Blended Semantic Browsing and Querying (BSBQ) provides domain specific search

(search based on relevant, domain specific attributes) and contextual browsing. The application involves crawling/extracting audio, video and text content from well over 250 sources (e.g. CNN website). This application was commercially deployed for a Web-audio company called *Voquette*.

69. Analytics and Knowledge Discovery. In the Passenger Threat Assessment application for national/homeland security, the knowledge base is populated from many public, licensed and proprietary knowledge sources. The resulting knowledge base has over one million instances. Periodic or continuous metadata extraction from tens of heterogeneous sources (150 files formats, HTML, XML feeds, dynamic Web sites, relational databases, etc) is also performed. When the appropriate computing infrastructure is used, the system is scalable to hundreds of sources, or about a million documents per day per server. A somewhat related business intelligence [IBMWF] application has demonstrated scalability by extracting metadata (albeit somewhat limited types of metadata with a significantly smaller ontology) from a billion pages.

70. Smart Search. Everything covered so far - emergent semantics, LSI, AI, NLP, personalization - are things that search engines already have experimented with, or will soon experiment with. Who does what first is what matters when it comes to competition in the search space. The ideal search engine is just something that brings as many of these techniques together as possible.

CONCLUSION

71. Facilities to put machine-understandable data on the Web are becoming a high priority for many communities. The Web can reach its full potential only if it becomes a place where data can be shared and processed by automated tools as well as by people. For the Web to scale, tomorrow's programs must be able to share and process data even when these programs have been designed totally independently. The Semantic Web is a vision: the idea of having data on the web defined and linked in a way that it can be used

by machines not just for display purposes, but for automation, integration and reuse of data across various applications.

72. The Semantic Web provides a common framework that allows data to be shared and reused across application, enterprise, and community boundaries. It is a collaborative effort led by W3C with participation from a large number of researchers and industrial partners. It is based on the Resource Description Framework (RDF), which integrates a variety of applications using XML for syntax and URIs for naming.

73. The development of the Semantic Web is well underway. This development is occurring in at least two areas: from the infrastructural, all-embracing, position as espoused by the W3C/MIT and other academically-focused organizations, and also in a more directed application-specific fashion by those using web technologies for electronic business.

BIBLIOGRAPHY

1. Deitel ,Deitel,Listfield,Nieto,Yaeger,Zlatkina.C# How to Program ,Chapter 18.

2. Digit Magazine,Aug issue, The Future of Search.

3. Tim Berners-Lee, James Handler, and Ora Lassia. The semantic web. *Scientific American*, 284(5):34–44, May 2001.

4. N. F. Noy, M. Sintek, S. Decker, M. Crubezy, R. W. Fergerson, and M. A. Musen. Creating Semantic Web Contents with Protege-2000. *IEEE Intelligent Systems* 16(2):60-71, 2001.

5. Semantic web community portal: http://www.semanticweb.org/.

6. The DARPA Agent Makeup Language (DAML) Program: http://www.daml.org/.

7. OntoWeb Consortium: http://www.ontoweb.org/.

8. Ontology Inference Layer (OIL): http://www.ontoknowledge.org/.

9. OntoBroker: http://ontobroker.aifb.uni-karlsruhe.de/.

DS COMMENTS

HOD COMMENTS

DEAN COMMENTS

DY COMDT AND CI COMMENTS