Workshop on
Digital Libraries: Theory and Practice
March,, 2003
DRTC, Bangalore

Paper: D

Semantic Web and Resource Description Framework (RDF)

Sneha Shukla
Documentation Research and Training Centre
Indian Statistical Institute
Bangalore- 560 059
email: *sneha@drtc.isibang.ac.in*

Abstract

Knowledge representation could be a powerful tool for search in digital collections. Semantic Web is commonly used knowledge representation technique in building Expert systems. It can be very well utilized in information retrieval in digital collections as well as Internet. Paper discusses the tools and techniques for knowledge representation using Semantic Web as well as its impact on the precision of search results in digital collections.

1. INTRODUCTION

Digital libraries are collections of digital objects. These objects are stored in a database or in hard drives but there may not be any arrangement at all. Situation is same when we look at the Internet, where there is no definite arrangement, no shape and no boundary. That is why searching for something over Internet often results in many irrelevant things. Discovery and accessing information on WWW is difficult and complex task due to semantic heterogeneities resulting from the different terminologies and conceptualization employed by various information providers and consumers. Proper representation of concepts of document and their relation could solve the problem to some extent and result in the chaotic web becoming a meaningful web or Semantic web.

Semantic literally means "The study of relationships between signs and symbols and what they represent". In semantic web, the concepts are defined in terms of their coordinate, subordinate and super-ordinate relationships. In library organization classification system forms a kind of semantic web.

2. CLASSIFICATION

We know Classification is nothing but "arrangement of entities in an order". When we talk about Library Classification, it is for arrangement of documents on shelf for easy retrieval of documents. But with digital collections in libraries, the concept of classification has changed from shelf arrangement to retrieval.

Traditionally, to classify, we use some kind of classification scheme like CC or DDC. These schemes give linear representation of Universe of Knowledge in a helpful sequence, still preserving the multi-faceted and hierarchical structure of Universe of Knowledge. Thus classification schemes are being used for Knowledge Representation defining the relative position of objects.

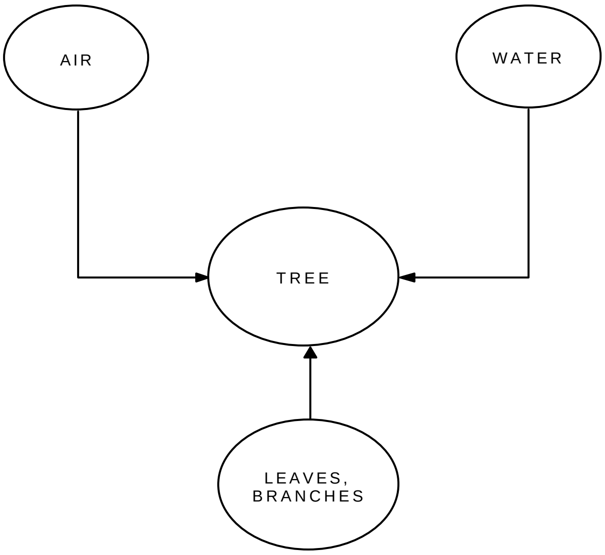
3. Ontology

The literal meaning of Ontology is "metaphysical study of nature of being and existence" (1). In other words, it is an explicit formal specification of how to represent the objects, concepts and other entities that are assumed to exist in some area of interest and the relationships that they hold among them, which is an implicit function of classification scheme. So, it holds the concept which forms the basis of developing a classification scheme. This particular concept is widely used in Knowledge representation with semantic web.

4. SEMANTIC NET

Take the example of a tree. The concept of a tree can be interpreted in relationship to many other things like air, water,etc.

In the above diagram we can see a kind of plexus of concepts which ultimately form a Web or net wholistically forming a semantic net. The idea of Semantic Net basically came from Aritificial Intelligence and is heavily used for knowledge representation in expert systems. An Expert system is not complete with its knowledge base as it has to have a mechanism of relating the concepts for inferencing (2).



There are various ways in which knowledge is represented based on the purpose. Main types are:

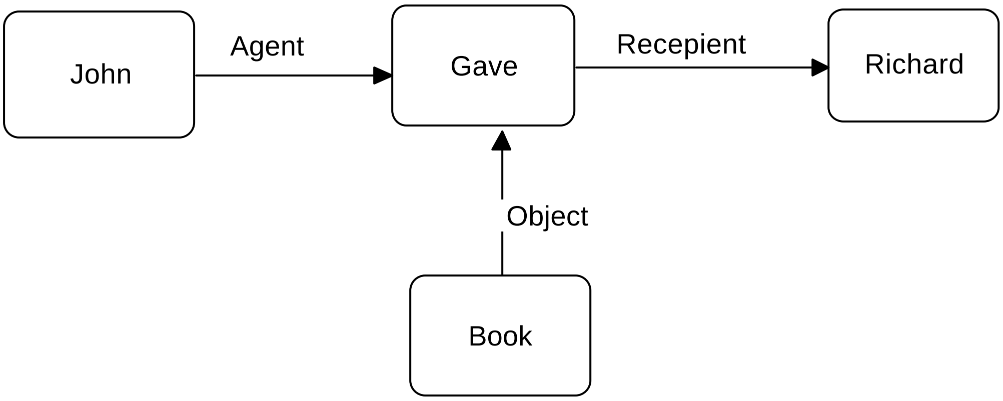
4.1. Semantic Nets

Semantic net mainly consists of two components:

a) points called 'nodes', and
b) 'arcs' which are the connecting links between two nodes.

Nodes can be a concept or an object or events. Arcs are used to represent the relationship existing between the two nodes. These can be employed for various purposes. This, generally, leads to a hierarchical structure. Arcs generally carry the notions like "belongs to" "is a" "has

e.g. John Gave Book to Richard.



4.2. Frame Based Knowledge Representation

Frames are much like semantic nets in structure but they basically stick to organization of facts in hierarchical way to eliminate repetition of common features by providing inheritance mechanism.
e.g., In order to define a person the basic characteristics can be his height, weight, profession, etc. These characteristics form different slots in the frame 'Person'.

4.3. Rule Based Knowledge Representation

Rules are employed to state the way in which inferencing has to be done.
e.g. *Rule:* IF it is cloudy and there is no wind THEN it will rain. That is when 'IF' portion of the rule is satisfied, then the 'THEN' portion is performed.
So, if the *facts* are

It is cloudy.
There is no wind.
Inference: It will rain.

The idea of semantic web came from the semantic net. However, Tim- Berners Lee introduced the term and idea of semantic web at WWW7 (Brisbane, 1997). At WWW8 (Toronto, 1998), he articulated the vision of a semantic web, whereby computers can have the capability of reasoning. They can interpret the things or the statements in the way we do.

5. REPRESENTATION OF ENTITIES (WEB DOCUMENTS OR DIGITAL DOCUMENTS) BY SEMANTIC NET

The basic problem while developing a Semantic Web is the representation of contents of a document. For this purpose markup languages are used. The oldest one is SGML (Standaridised Generalised Markup Language). It was developed by the International Organization for Standardization (ISO) as a method for describing documents in a way that makes it easy to move them from one platform to another. HTML is another language which is basically known as formatting language and to a limited extent used to structure the data.

5.1. XML

While HTML allows us to visualize the information on the web, it doesn't provide much capability to describe the information in ways that facilitate the use of software programs to find or interpret the information. We can use eXtensible Markup Language where we can define our own tags, unlike in HTML where we have to stick to certain specified tags. (3)
e.g. To define a resource "book" titled "Prolegomena to Library Classification" authored by "S. R. Ranganathan", can be represented in a XML document as

```
<book>
<title> Prolegomena to Library Classification</title>
<author>S. R. Ranganathan</author>
</book>
```

But with XML, non-standardization was basic problem as anyone can have his/her own tag. Metadata Schemas provide a standard set of terms to define a resource belonging to a particular domain of knowledge.

5.2. Metadata

Metadata is the "data about data". In case of digital libraries, we have schemas like DC and RDF which have certain set of elements to describe the document.

5.2.1 DC: Developed in Dublin, Ohio in 1995 under Dublin Core Metadata Initiative. It is a set of 15 elements intended to cataloguing and searching the electronic resource.

5.2.2 SHOE: SHOE is an SGML/XML 'HTML-based' knowledge representation language -- "a superset of HTML which adds the tags necessary to embed arbitrary semantic data into web pages. SHOE tags are divided into two categories. First, there are tags for constructing ontologies. SHOE ontologies are sets of rules which define what kinds of assertions SHOE documents can make and what these assertions mean. Secondly, there are tags for annotating web documents to subscribe to one or more ontologies, declare data entities, and make assertions about those entities under the rules proscribed by the ontologies.

5.2.3 DAML: DAML stands for the DARPA Agent Markup Language, which is a project being funded by the US Defense Advanced Research Projects Agency -- the same organization that funded much of the original work on the Internet (which was then called the ARPAnet). DARPA is developing not only the language, but many tools and applications that promote its use. (6)

5.2.4 OIL: OIL stands for the Ontology Interchange Language and was developed by a number of researchers, primarily a group funded by the European Union's Information Society Technologies Program. This uses constructs from Frame-Based AI for more sophisticated classification.

5.2.5 DAML+OIL*:* The most recent, and currently most used, is a language called DAML+OIL which was developed under joint sponsorship of US and European government agencies. DAML+OIL is based on the Resource Description Framework (RDF) which is in turn based on the eXtensible Markup Langauge (XML). It has the best features of SHOE, DAML, OIL and several other markup approaches.

5.2.6 Resource Description Framework (RDF): is a specification language incorporating aspects from knowledge representation models (e.g. Semantic Nets), and database schema definition languages. It is a simple language of restricted expressive power. RDF(S) define classes and properties which can be used to describe a resource.

5.3. Resource Description Framework

To represent the knowledge in a web page there is a document description framework developed known as ***RDF (Resource Description Framework).***

5.3.1 History: The development of RDF started with the initiation of PICS (Platform for Internet Content Selection) project in 1995. PICS was a rating mechanism about the contents of web pages. The idea was to filter the unwanted set of web pages, which contain foul language, pornographic material, violence etc. Once the project was initiated it was found that it can be used for describing the content of web page and could be made to represent content understandable by machines. The extension of PICS project was PICSNG (PICS Next Generation), which was later called as RDF.

5.3.2 Uses Of RDF: RDF can be used in a variety of application (2) areas;

1) In *resource discovery* to provide better search engine capabilities,
2) In *cataloging* for describing the content and content relationships available at a particular Web site, page, or digital library, by *intelligent software agents* to facilitate knowledge sharing and exchange,
3) In *content rating*, in describing *collections of pages* that represents a single logical "document",
4) For describing *intellectual property rights* of Web pages, and
5) For expressing the *privacy preferences* of a user as well as the *privacy policies* of a Web site. RDF with *digital signatures* will be key to building the "Web of Trust" for electronic commerce, collaboration, and other applications. (3)

The broad goal of RDF is to define a mechanism for describing resources which is domain independent. In other words, it should be neutral.

5.3.3 Basic RDF Model: The basic data model consists of three object types:

Resources: A resource can be defined as any existing entity having some property or attributes. In other words,

> "Any entity which has to be described is known as Resource which is equivalent to *Subject* in normal English grammar." (3)

A resource can be a book, an entire web page or a part of it or it can be an entire web site.

Properties: Any characteristic of Resource or its attribute which is used for the description of the same is known as Property, which is equivalent to *Predicate* in normal English grammar. e.g., a web page can be recognized by 'Title' or a man can be recognized b both are attributes for recognition of resource 'webpage' and 'person' respectively.

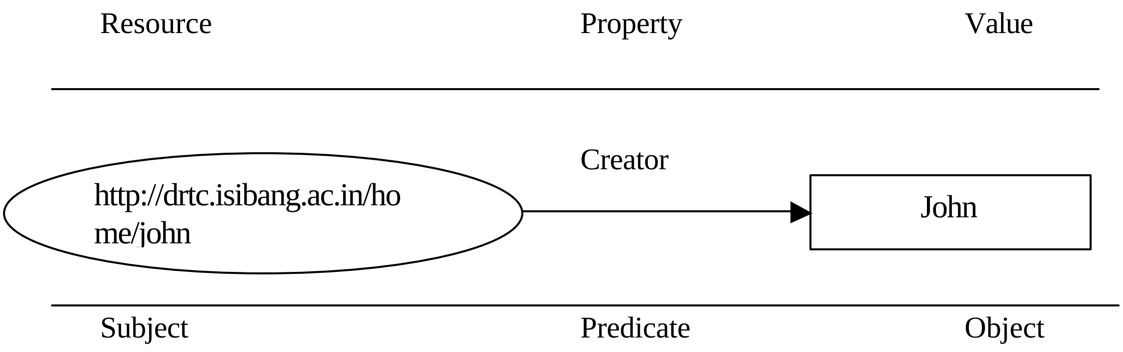
Values: A Property must have a value which is equivalent to *Object* in normal English grammar. It can be another resource or it can be a literal.

Statements: A specific resource together with a named property plus the value of that property for that resource is an RDF *statement*. These three individual parts of a statement are called, respectively, the *subject*, the *predicate*, and the *object*.
eg. *John is the creator of the resource http://drtc.isibang.ac.in/home/john*

This sentence has the following parts:

Subject (Resourc)	http://drtc.isibang.ac.in/home/john
Predicate (Property)	Creator
Object (literal)	"John"

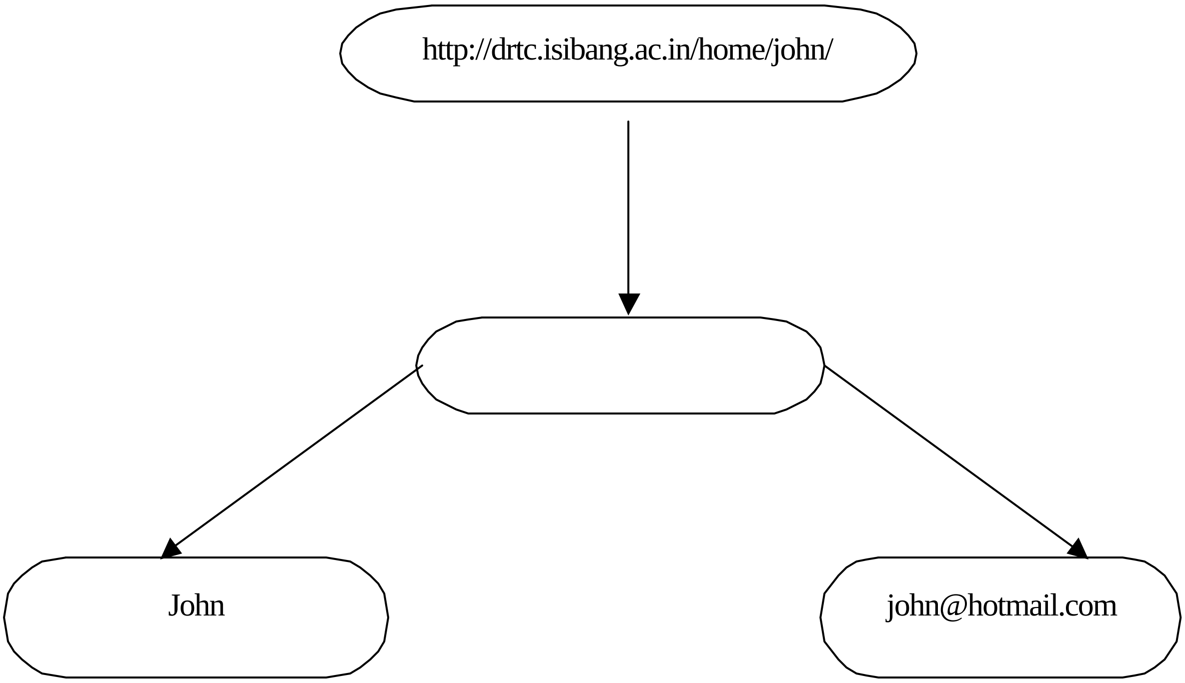
The above example can be represented graphically also (also called "nodes and arcs diagrams"). In this diagram, the nodes (drawn as ovals) represent resources and arcs represent named properties. Nodes that represent string literals will be drawn as rectangles. The sentence above would thus be represented in the following diagram



The direction of the arrow is important. The arc always starts at the subject and points to the object of the statement. The simple diagram above may also be read "*http://drtc.isibang.ac.in/home/john has creator John*", or in general "*<subject>has<predicate><object>*".

Now, consider the case that we want to say something more about the characteristics of the creator of this resource. For example,
The individual whose name is John, with an email <john@hotmail.com>, is the creator of http://drtc.isibang.ac.in/home/ *john*

The intention of this sentence is to make the value of the Creator property a structured entity. In RDF such an entity is represented as another resource. The sentence above does not give a name to that resource; it is anonymous, so in the diagram below we represent it with an empty oval:

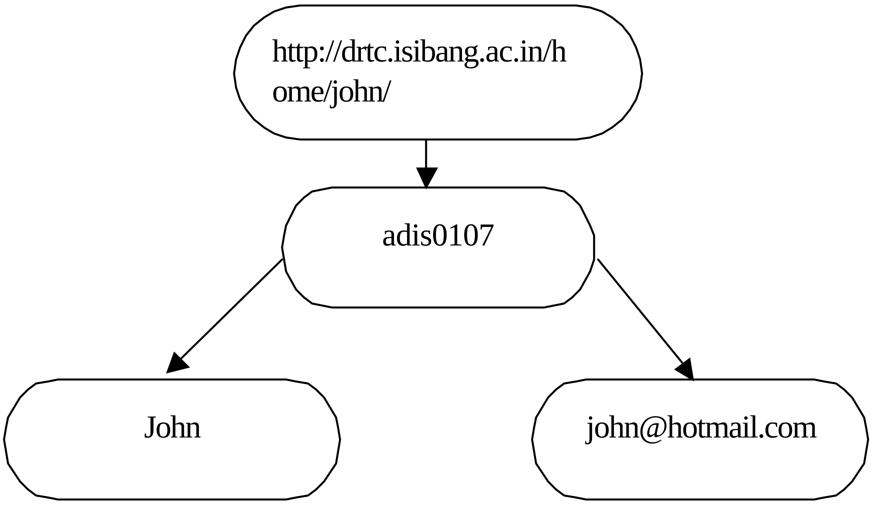


This diagram could be read "http://drtc.isibang.ac.in/home/john" has creator and the author has name John and email john@hotmail.com".

The structured entity of the previous example can also be assigned a unique identifier. The choice of identifier is made by the application database designer. To continue the example, imagine that a person's id is used as the unique identifier for a "person" resource. Lets take another example

The individual referred to by student id adis0107 is named John and has the email address john@hotmail.com. The resource http://drtc.isibang.ac.in/home/john was created by this individual.

The RDF model for these sentences is:



The RDF data model provides an abstract, conceptual framework for defining and using metadata. A concrete syntax is also needed for the purposes of creating and exchanging this metadata. This specification of RDF uses the eXtensible Markup Language [XML] encoding as its interchange syntax. RDF also requires the XML namespace facility to precisely associate each property with the schema that defines the property.

5.3.4 RDF Syntax

This specification defines two XML syntaxes for encoding an RDF data model instance. The *serialization syntax* expresses the full capabilities of the data model in a very regular fashion. The *abbreviated syntax* includes additional constructs that provide a more compact form to represent a subset of the data model. RDF interpreters are expected to implement both the full serialization syntax and the abbreviated syntax. Consequently, metadata authors are free to mix the two.

The file starts by declaring that it is written in RDF and defines the number of namespaces it uses.

```
<rdf : RDF>
   <xmlns:
     <rdf : Description   ID= “any legal XML name symbol(IDsymbol)if the resource does
not yet exist” | about= “URI Reference”> (property of Element)
        <(propName: :=Qname) Nsprefix : name> value </propName> |
        <propName (resourceAttr) resource : “URI-reference” />
     </rdf : Description>
<rdf: RDF>
```

A Generalized RDF Syntax

For the example sentence given earlier
Sneha is the creator of the resource http://http.drtc.isibang.ac.in/home/sneha.
is represented in RDF/XML as:

```
<rdf:RDF
      xmlns:rdf ="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
      xmlns:s="http://uri_of_s_standard#" >
  <rdf : Description about="http://drtc.isibang.ac.in/home/sneha”>
    <s : Creator>Sneha</s : Creator>
   <s:Email>sneha@hotmail.com</s:Email>
  </rdf : Description>
</rdf : RDF>
```

For more examples, see **Appendix- I and Appendix- II**

6. CHALLENGES BEFORE US

6.1. Ontology Availability, Development And Evolution

a) Developing an ontology which can be used by all domains,

b) Methodological and technological support for it’s development like Ontology alignment and mapping, Ontology integration, Ontology translation tools and Ontology reengineering tools if existing ontologies are going to be used, and

c) Evolution of ontologies and their relation to already annotated data. Configuration management tools are necessary to keep control of the versions of each ontology as well as the interdependencies between them and annotations.

6.2. Scalability of Semantic Web Content

Once we have Semantic Web, the next problem to reckon with is managing it in a scalable manner.

a) Storage and organization of Semantic Web pages: In a Semantic Web environment, all the documents exist in relation with each other and are interlinked to form WWW. This interlinking is done by means of hyperlinks. But these hyperlinks do not fully exploit the underlying semantics of Semantic Web pages. So there is a need to develop “Semantic Indexes” to group Semantic documents based on particular topics.

b) Easy Finding of Information: This can be achieved by coordination among Semantic Indexes. A peer to peer kind of networking is suggested for this purpose. Indexes are considered as active agents that know which topics to handle and can pass the topics to neighbour indexes if not found in their indexes.

6.3. Multilinguality

We know that a number of documents are available in languages other than English. Multilinguality plays an increasing role at following levels:

a) at the level of ontologies: Ontology builders may want to use their native language for the development of the ontologies in which annotations will be based. Since not all the users will be ontology builders this level has got low priority compared to other levels.

b) at the level of annotations: Annotation of content can be performed in various languages. Since more users will go for annotating their content rather than developing ontologies. A good system is needed to allow the users to annotate their content in their native language.

c) at the level of user interface : users would like to access the relevant content in their native language irrespective of the source language in which annotations are presented. The approach of “personalisation of information” should be explored.

7. IMPLICATIONS OF HAVING SEMANTIC WEB

1. One of the foremost utilities of Semantic Web is Information Retrieval. We can describe our documents for better search. Hence, if we give a search for Mr. Cook it will not bring the information related to cooking or cookies. Thus information search becomes more efficient in a semantic web environment as well as the goal of global knowledge sharing can be achieved.
2. Information is given well defined meaning.
3. There is a full utilization of web potential. If we have semantic web we can envision that some day we can sit in our offices and can operate our home appliances. The web can be applicable in our day to day processes.
4. Ability to assess the trustworthiness of the information. As, whatever information will occur on web, it will have some context. Applications that will be needed to process this will evaluate the authenticity of information.

Everything or any entity defined in a semantic web have an identity whether it is people, places or things, all have Identifiers on web.

1. This narrows down the gap between the way we interpret the things and the way computers do.
2. More cost-effective for people to effectively record their knowledge.
3. Economic Aspect: Web Based Services can be made more effective using Semantic web. Many projects are going on in this direction e.g.

    a) UDDI: The Universal Description, Discovery and Integration (UDDI) project creates a framework for describing services, discovering businesses, and integrating business services using the Internet.

    b) WSDL: The Web Service Description Language is an XML format for describing interfaces to business services registered with a UDDI database.

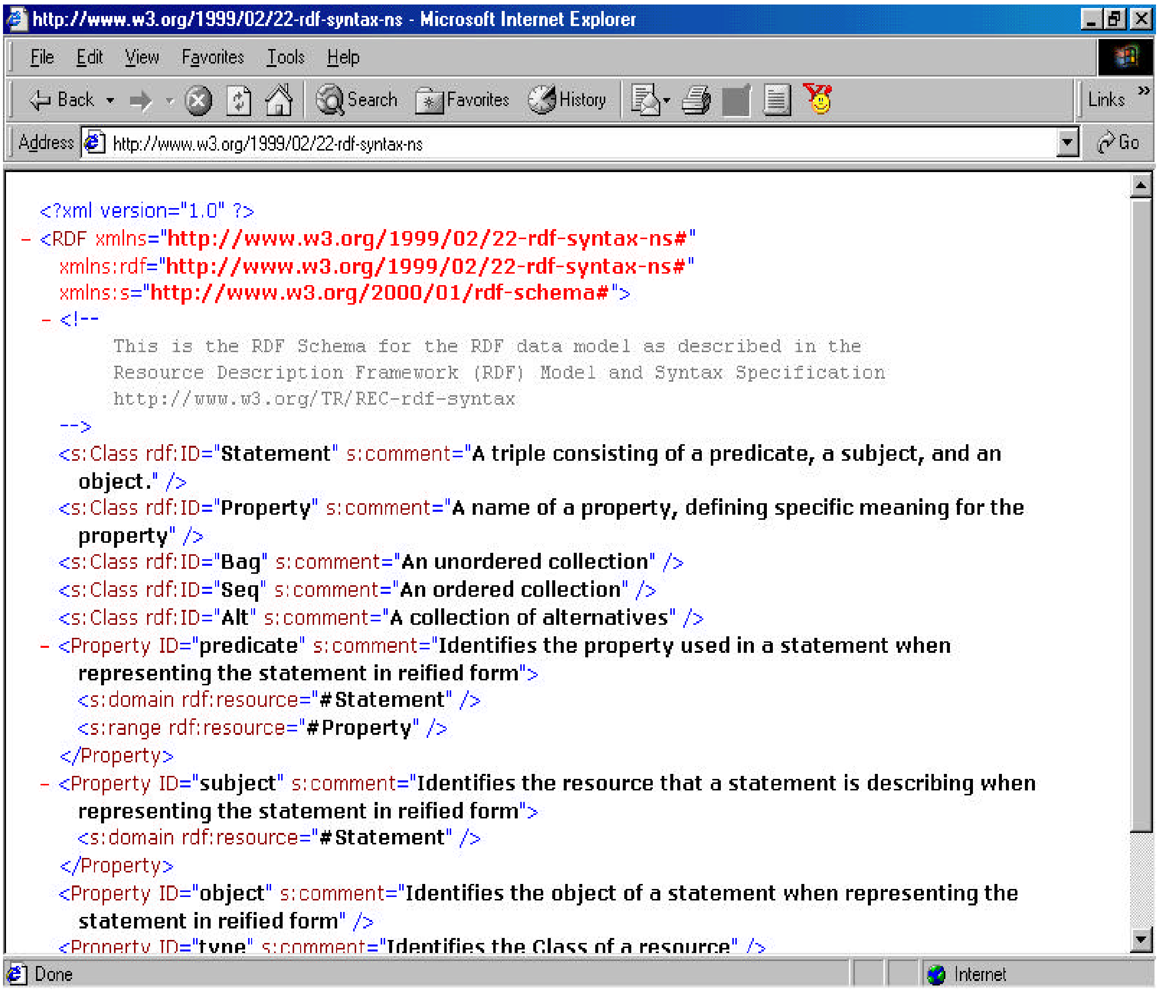
8. CONCLUSION

Semantic Web means different to different people. To some, it is for easy retrieval of the information they need having proper authenticity in a less span of time. Others visualize it as a way for global knowledge sharing and enabling the computers and the people to work in cooperation. Whatever is the idea, it is a grand vision to have Semantic web. However, there is still a long way to go and many obstacles have to be overcome. The possibilities are endless and even than if we will not able to achieve it the journey itself be a reward.

9. REFERENCES

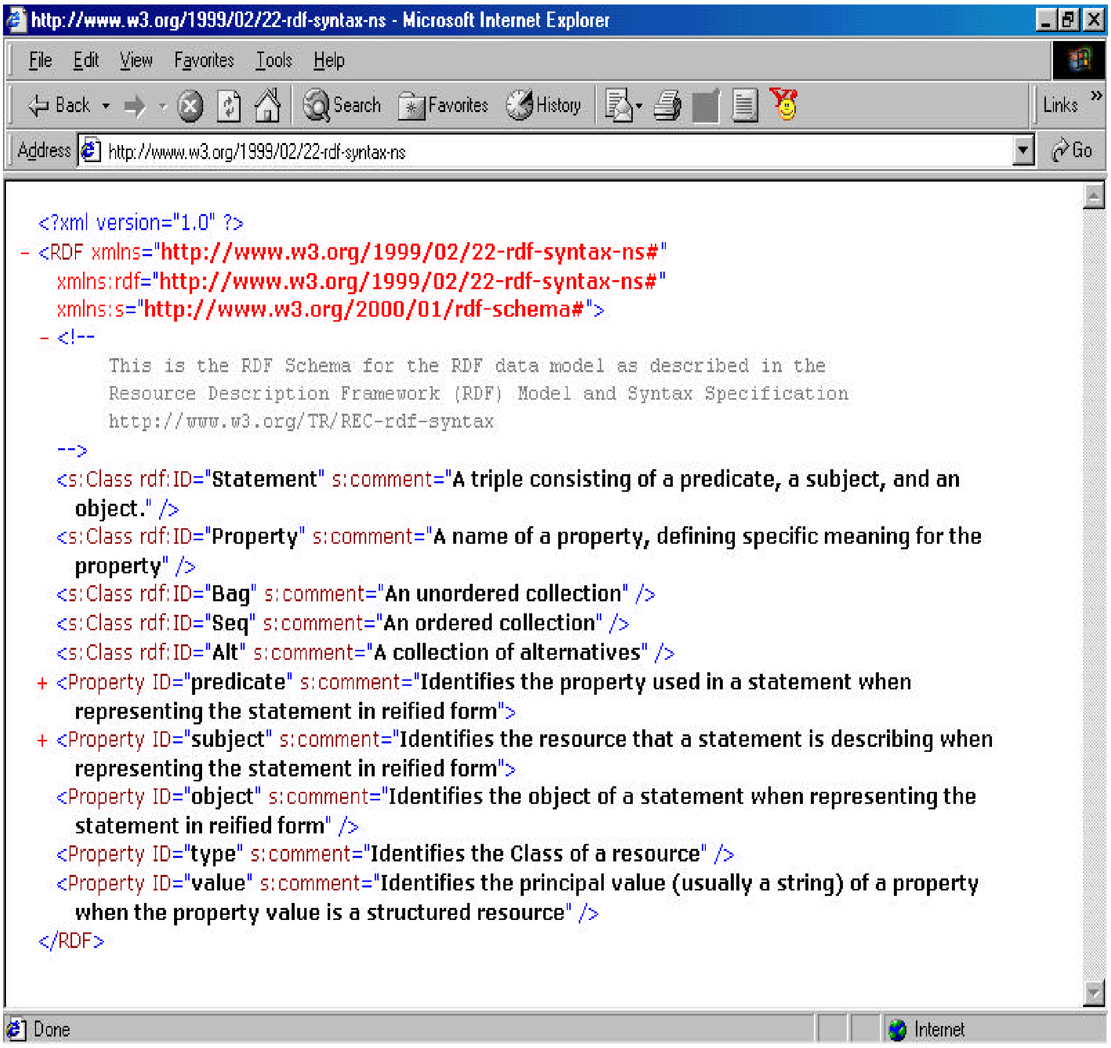
1. Amman, B. et al(2002). Integrating ontologies and thesauri. International Journal on Digital Libraries, *3*(3), 221- 236.
2. Resource Description Framework (RDF) Model and Syntax Specification. from http://www.w3.org/TR/1999/REC-rdf-syntax-19990222/
3. Electronic Commerce: A Killer (Application) for Semantic web by Dieter Fensel. from http://www.cs.vu.nl/~dieter/ftp/slides/eunsf.pdf
4. The DARPA Agent Markup Language Homepage. from http://www.daml.org
5. Backett, D. (2002). The design and implementation of the Redland RDF application framework. Computers Network, *39*, 577-588.
6. Online resource for markup language technologies. from http://xml.coverpages.org
7. Ding, Y. et al(2002). The semantic web: yet another hip? *Data Information and Knowledge, 41*, 205-227.

Appendix- I


http://www.w3.org/1999/02/22-rdf-syntax-ns - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites History
Links
Address http://www.w3.org/1999/02/22-rdf-syntax-ns
Go
Done
Internet

```
  <?xml version="1.0" ?>
- <RDF xmlns="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:s="http://www.w3.org/2000/01/rdf-schema#">
  - <!--
         This is the RDF Schema for the RDF data model as described in the
         Resource Description Framework (RDF) Model and Syntax Specification
         http://www.w3.org/TR/REC-rdf-syntax
    -->
    <s:Class rdf:ID="Statement" s:comment="A triple consisting of a predicate, a subject, and an
      object." />
    <s:Class rdf:ID="Property" s:comment="A name of a property, defining specific meaning for the
      property" />
    <s:Class rdf:ID="Bag" s:comment="An unordered collection" />
    <s:Class rdf:ID="Seq" s:comment="An ordered collection" />
    <s:Class rdf:ID="Alt" s:comment="A collection of alternatives" />
  - <Property ID="predicate" s:comment="Identifies the property used in a statement when
      representing the statement in reified form">
      <s:domain rdf:resource="#Statement" />
      <s:range rdf:resource="#Property" />
    </Property>
  - <Property ID="subject" s:comment="Identifies the resource that a statement is describing when
      representing the statement in reified form">
      <s:domain rdf:resource="#Statement" />
    </Property>
    <Property ID="object" s:comment="Identifies the object of a statement when representing the
      statement in reified form" />
    <Property ID="type" s:comment="Identifies the Class of a resource" />
```
]]>

Appendix- II


http://www.w3.org/1999/02/22-rdf-syntax-ns - Microsoft Internet Explorer
File
Edit
View
Favorites
Tools
Help
Back
Search
Favorites
History
Links
Address
http://www.w3.org/1999/02/22-rdf-syntax-ns
Go
Done
Internet

```
  <?xml version="1.0" ?>
- <RDF xmlns="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:s="http://www.w3.org/2000/01/rdf-schema#">
  - <!--
         This is the RDF Schema for the RDF data model as described in the
         Resource Description Framework (RDF) Model and Syntax Specification
         http://www.w3.org/TR/REC-rdf-syntax
    -->
    <s:Class rdf:ID="Statement" s:comment="A triple consisting of a predicate, a subject, and an
      object." />
    <s:Class rdf:ID="Property" s:comment="A name of a property, defining specific meaning for the
      property" />
    <s:Class rdf:ID="Bag" s:comment="An unordered collection" />
    <s:Class rdf:ID="Seq" s:comment="An ordered collection" />
    <s:Class rdf:ID="Alt" s:comment="A collection of alternatives" />
  + <Property ID="predicate" s:comment="Identifies the property used in a statement when
      representing the statement in reified form">
  + <Property ID="subject" s:comment="Identifies the resource that a statement is describing when
      representing the statement in reified form">
    <Property ID="object" s:comment="Identifies the object of a statement when representing the
      statement in reified form" />
    <Property ID="type" s:comment="Identifies the Class of a resource" />
    <Property ID="value" s:comment="Identifies the principal value (usually a string) of a property
      when the property value is a structured resource" />
  </RDF>
```