

# Information Organisation: Information Retrieval

Jagdish S. Gangolly,  
Department of Informatics,  
SUNY Albany

November 1, 2009

- ▶ Motivation
- ▶ Semantic Web Technologies (Layered Approach)
- ▶ XML Language
- ▶ Resource Description Framework (RDF)

# Motivation

- ▶ Problems with present day information retrieval
- ▶ Problems with present day documents: Document tags are mostly of the formatting and structure kind. The meaning of the information in the documents are not tagged (except for the structured data when tagged using xml-schema)
- ▶ Need for machine interpretation of search results
- ▶ Need for machines to interchange information/documents without human interference
- ▶ Need to support inter-personal collaboration

# Problems with Present Day Information Retrieval

- ▶ High recall, low precision
- ▶ Search results sensitive to vocabulary
- ▶ The basic unit for retrieval is a document. Query results can straddle more than one document

- ▶ Explicit Metadata
- ▶ Ontologies
- ▶ Logic & Reasoning
- ▶ Software Agents

# Layered Approach to Semantic Web

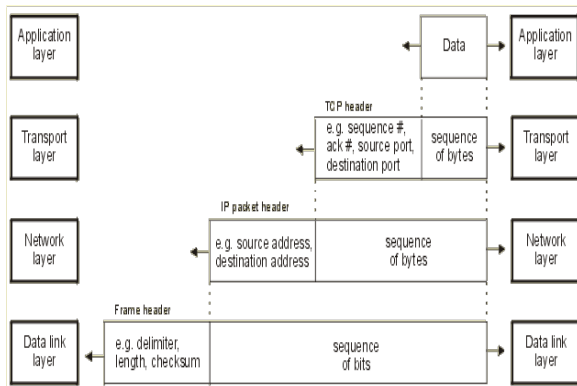


Figure: Information Model Interoperability Reference Model

Source: <http://infolab.stanford.edu/~melnik/pub/sw00/>

# Layered Approach to Semantic Web

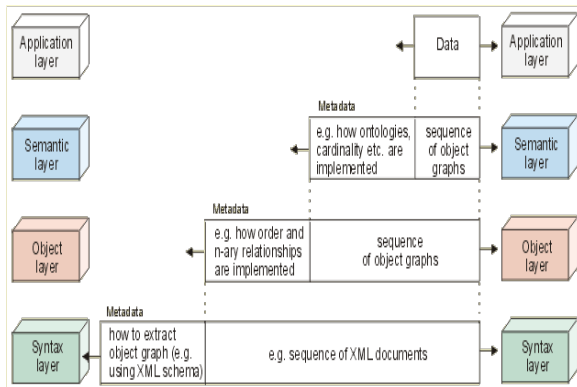


Figure: Data Modeling Layers for the Semantic Web

Source: <http://infolab.stanford.edu/~melnik/pub/sw00/>

# Layered Approach to Semantic Web

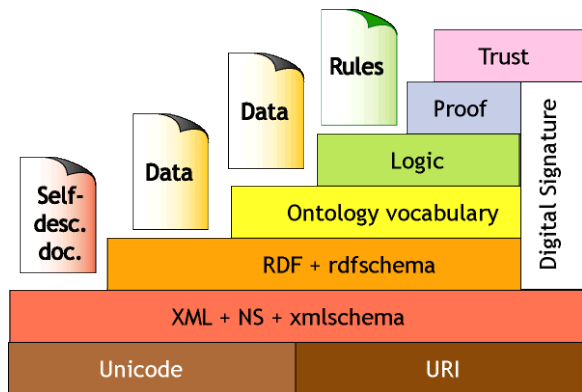


Figure: Semantic Web Layered Architecture

Source: <http://www.w3.org/2000/Talks/1206-xml2k-tbl/slide10-0.html>

- ▶ HTML has mostly two types of tags
  - ▶ formatting tags such as *b*, *i*, *u*, *center*, *br*, ...
  - ▶ structure tags such as *head*, *title*, *body*, *div*, ...
- ▶ XML allows one to extend the tagset by defining additional tags, hence the name *Extensible* Markup Language. It therefore allows one to put impose (or describe) some semantic structure on (or of) the documents.

**An Example:** `<?xmlversion = "1.0"? >`

`< catalog >`

`< bookid = " bk101" >`

`< author > Gambardella, Matthew < /author >`

`< title > XMLDeveloper'sGuide < /title >`

`< genre > Computer < /genre >`

`< price > 44.95 < /price >`

`< publish_date > 2000 - 10 - 01 < /publish_date >`

`< description > Anin - depthlookatcreatingapplicationswithXML.`

`< /description >< /book >< catalog >`

- ▶ Prolog: XML Declaration and reference to external Document Type Definition (DTD)

## **An Example:**

```
<?xmlversion = "1.0" encoding = " UTF - 8"?standalone = " no" >  
<!DOCTYPEgreetingSYSTEM" hello.dtd" >
```

or,

```
<?xmlversion = "1.0" encoding = " UTF - 8"?standalone = " yes" >  
<!DOCTYPEgreeting[  
<!ELEMENTgreeting(#PCDATA) >] >
```

- ▶ Elements and Attributes
- ▶ Comments
- ▶ Processing Instructions
- ▶ Document Type Definitions and XML Schema
- ▶ Namespaces
- ▶ ...

# Resource Description Framework (RDF)

- ▶ While XML provides the basic infrastructure beyond HTML for documents, it does not facilitate description of semantic relationships between domain concepts.
- ▶ RDF is basically a data model whose basic building block is a resource-attribute-value triple
- ▶ RDF Schema provides the syntax for the description of the vocabulary for the development of RDF for any application.

- ▶ Resources, uniquely identified by a URI (Uniform Resource Identifier). A URI can be a URL.
- ▶ Properties, also identified by URIs, are relations between resources
- ▶ Statements assert properties of resources. A statement is a triple  $(x, P, y)$  where  $x$  and  $y$  are objects and  $P$  relates  $x$  and  $y$ . RDF offers only *binary predicates*.
- ▶ Graph corresponding to RDF statements is a semantic network.
- ▶

# RDF: An Example

"there is a Person identified by <http://www.w3.org/People/EM/contact#me>, whose name is Eric Miller, whose email address is [em@w3.org](mailto:em@w3.org), and whose title is Dr."

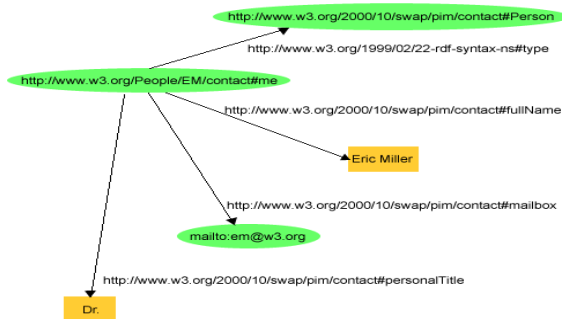


Figure: An RDF Graph Describing Eric Miller

Source: <http://www.w3.org/TR/REC-rdf-syntax/#figure1>

# RDF: An Example

RDF uses URIs for

- ▶ individuals, e.g., Eric Miller, identified by <http://www.w3.org/People/EM/contact#me>
- ▶ kinds of things, e.g., Person, identified by <http://www.w3.org/2000/10/swap/pim/contact#Person>
- ▶ properties of those things, e.g., mailbox, identified by <http://www.w3.org/2000/10/swap/pim/contact#mailbox>
- ▶ values of those properties, e.g. <mailto:em@w3.org> as the value of the mailbox property (RDF also uses character strings such as "Eric Miller", and values from other datatypes such as integers and dates, as the values of properties)

**Source:** <http://www.w3.org/TR/REC-rdf-syntax/>

# RDF: An Example

```
<?xmlversion = "1.0"? >
```

```
< rdf : RDFxmlns : rdf = "http : //www.w3.org/1999/02/22 - rdf - syntax - ns#"
xmlns : contact = "http : //www.w3.org/2000/10/swap/pim/contact#" >
```

```
< contact : Personrdf : about = "http : //www.w3.org/People/EM/contact#me" >
```

```
< contact : fullName > EricMiller < /contact : fullName >
```

```
< contact : mailboxrdf : resource = "mailto : em@w3.org" / >
```

```
< contact : personalTitle > Dr. < /contact : personalTitle >
```

```
< /contact : Person >< /rdf : RDF >
```

**Source:** <http://www.w3.org/TR/REC-rdf-syntax/>