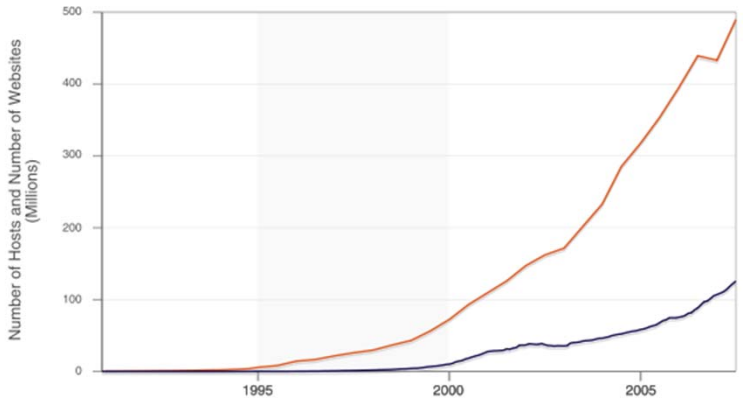
Web and the Internet

The Web

- ✓ World's largest application platform
- ✓ Giant virtual disk
- ✓ Giant hyperlinked document
- ✓ The growth is staggering
- ✓ Some statistics...

The Web: The Growth of the Number of Internet Hosts and Websites

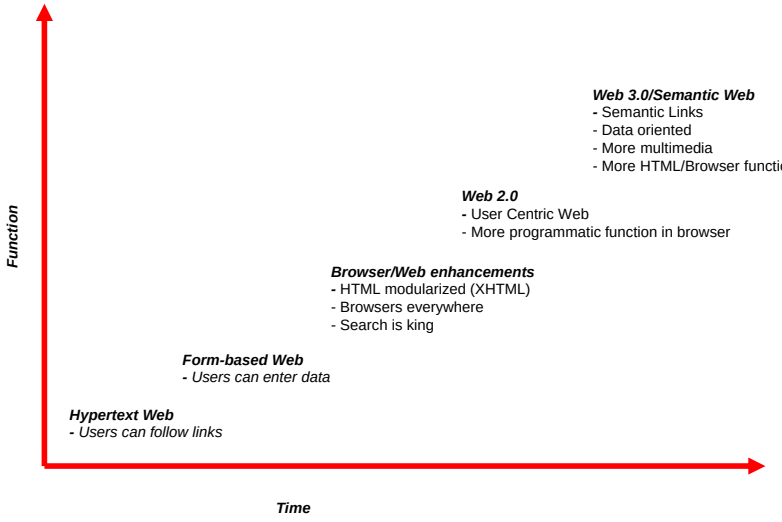


*Sources: Internet Systems Consortium (http://www.isc.org/index.pl...) and netcraft and Hobbes Internet Timeline (http://www.zakon.org/robert/internet/timeline)

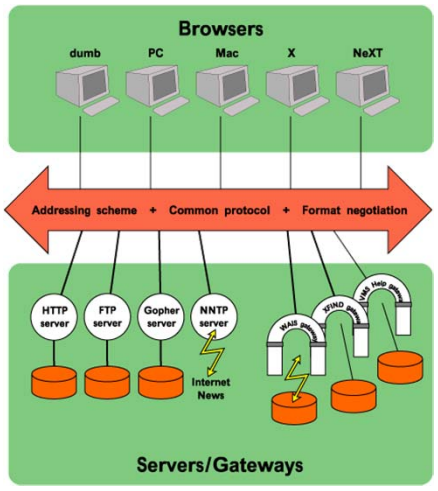
Web Client/Server: Portable Protocols

- ✓ KISS
- ✓ Platform Independent
- ✓ Internet as backbone
  - Global and Private
  - 100,000's of interconnected networks
  - Subsumes many early Internet protocols
    - SMTP, Telnet, FTP, NNTP, Gopher
  - Adds new Protocols/Standards
    - HTTP, HTML, XHTML, URL
- ✓ URL Global naming
- ✓ HTTP RPC-Like protocol
- ✓ HTML documents
  - Tagged Text Documents
  - Made available from Web servers
- ✓ Web browsers as universal clients
  - Not just PCs—mobile phones, PDAs, kiosks, etc.

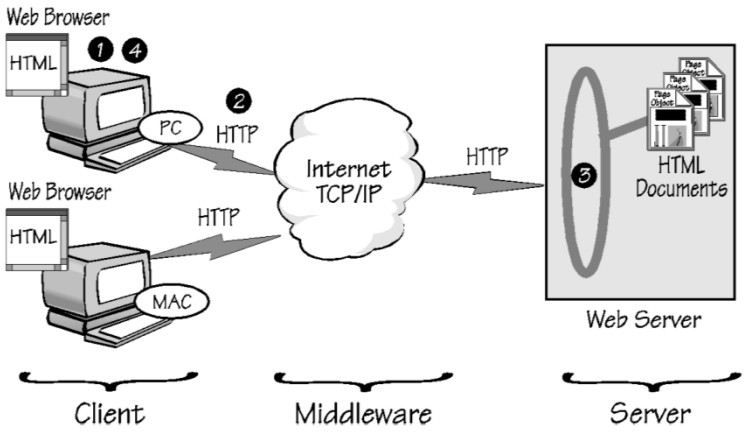
Web Evolution…



Web Architecture at the 50,000 foot level



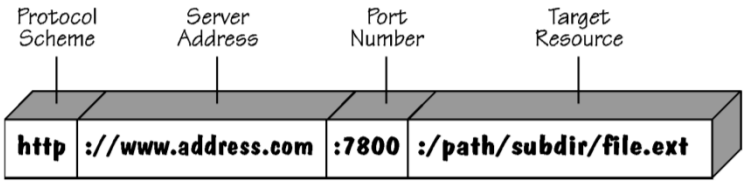
A Web Client/Server Interaction



Simple Web Client/Server Interaction

- ✓ 1. Select target URL
- ✓ 2. Browser send HTTP request to server
- ✓ 3. Server processes request
  - Well-known port 80 for HTTP
  - send back requested HTML file; close connection
- ✓ 4. Browser interprets HTML commands
  - If HTML type, displays page
  - Otherwise, launches helper

The URL Structure



- ✓ URI: Uniform Resource Identifier—locates a "point of content" on the Web.
- ✓ URL: Uniform Resource Locator—subset of URIs that locates files on the Internet
  - If you have the URL and an appropriate protocol you can retrieve file

URL (Uniform Resource Locator)

- ✓ Naming scheme + how to get to resource
- ✓ Protocol scheme
  - HTTP, Gopher, News, FTP, WAIS, Mailto, nntp, Telnet
  - HTTP is native Web protocol
- ✓ Server name
  - Internet host domain name or raw IP address
- ✓ Port number (or default)
  - HTTP = 80; Gopher = 70; FTP = 21
- ✓ Path to resource

HTML: Hyper Text Markup Language

- ASCII file
  - text + tags
- Tags are commands
  - Tag-pairs--command/inverse-command
- Document structure
  - Header/Body
- Rooted in ISO SGML (Standard Generalized Markup Language)
  - DTD (Document Type Definition)
  - Extensions for hyper-linking

HTML Versions

- HTML 2.0
  - Developed upon current practice in 1994 by the ITEF (which closed in 1996)
- HTML 3.2
  - First Version from the W3C, Jan 97
  - Released in 1996
  - Added tables, applets, text-flow around images, super/subscripts
  - Backward compatible with 2.0
- HTML 4.0
  - First released in Dec 97, Second release April 98
  - Adds more more multimedia options, scripting languages, style sheets, better printing facilities, and documents that are more accessible to users with disabilities
- HTML5
  - Next HTML Revision
  - Currently in “Last Call” status in the W3C Web Hypertext Application Technology Working Group (WHATWG)
  - Goal is to reduce the use of proprietary Rich Internet Application (RIA) application plug-ins such as Adobe Flash, Microsoft Silverlight, and Sun JavaFX.

Problems with HTML

- ✓HTML Challenges
  - Sloppy markup practices
  - New browsers
    - Digital TVs, Handhelds, phones, cars
  - Needed markup subset for simpler clients
  - Needed extended markup for richer clients
  - Needed to combine markup with other tag sets
    - Math, Vector graphics, E-commerce, metadata

Solution: XHTML – The Next Generation Markup Language

- ✓XHTML: eXtensible Hypertext Markup Language
  - A reformulation of HTML in XML with namespaces for HTML 4.0 strict, transitional and frameset DTDs
  - Modularizes HTML for subsetting/combining with other tag-sets
  - Document Profiles provide basis for interoperability guarantees
  - Next generation forms features offering improved match to database and workflow applications

XHTML

- A family of document types (called XHTML Family) which collectively form a markup language
- Introduced in Three Steps
  1. XHTML 1.0, Jan 2000
     - Defines HTML 4 using XML 1.0
     - Three DTDs: Strict, Transistional, Frameset
  2. XHTML 1.1: April 2001
     - Modularization of HTML
  3. XHTML 2.0
     - Legacy cleanup, improved structure, Xforms, XML events, ...
- XHTML5 is being defined alongside HTML5 in the HTML5 draft specification.

XHTML 1.0

- Most differences come from differences between SGML and XML
  - Documents must be well-formed
    - must have matching <...> </...>
  - Element and attribute names must be in lower case
  - Elements and attributes must be written in full
  - etc.

XHTML 1.1

- Four core Modules
  - Structure Module (html, head, body, title)
  - Text Module (abbr, acronym, address, blockquote, br, cite, code, dfn, div, em, h1-h6, kbd, p, pre, q, samp, span, strong, var)
  - Hypertext Module (a)
  - List Module (dl, dt, dd, ol, ul, li)

XHTML 1.1 (cont)

- Other XHTML 1.1 Modules
  - **Text Extension Modules**: Presentation Module, Edit, Bi-directional Text
  - **Forms Modules**: Basic Forms, Forms
  - **Table Modules**: Basic Tables, Tables
  - **Miscellaneous Modules**: Image, Client-side Image Map, Server-side Image Map, Object, Frames, Target, Iframe, Intrinsic Events, Metainformation, Scripting, Style Sheet, Style Attribute, Link, Base

XHTML 1.1 (cont)

- **XHTML Host Language Document Type**
  - A document type which uses XHTML as a host language
  - Must include Structure, Text, Hypertext, and List modules
  - Other Document types
    - Defined by W3c
      - XHTML Basic, XHTML 1.1, XHTML+MathML+SVG etc.
    - Defined by other organizations
      - WML 2.0, XHTML-Print etc.

XHTML 1.1 (cont)

- **XHTML Integration Set Document Type**
  - A document type which integrates XHTML modules into another host language
  - At a minimum, a document type MUST include Text, Hypertext and List Modules (Structure Module is not required)

Moving to XML

- ✓ HTML was created as an application of SGML - the Standard Generalized Markup Language (ISO 8879:1986)
- ✓ XML is a descendant of SGML, which is easier to implement
- ✓ XML requires you to:
  - make tags case-sensitive
  - include end tags e.g. </p> and
  - add a / to empty tags, e.g.
 and <hr />
  - quote all attribute values, e.g.
- ✓ These make it practical to parse well-formed XML without *a priori* knowledge of the tags
- ✓ XHTML uses *lower-case* for tags and attributes
- ✓ Old browsers can render XHTML 1.0 provided you follow simple guidelines

For More Information on HTML/XHTML…

- ✓ see http://www.w3.org/MarkUp/

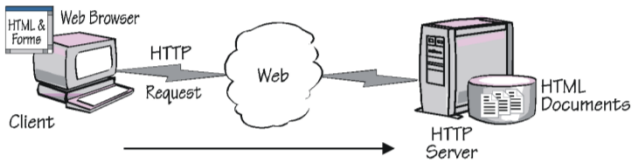
HTTP Standard

- ✓ In use since 1990
- ✓ HTTP 0.9
  - The original HTTP, defined in 1991
- ✓ HTTP/1.0
  - Early versions in 1992
  - Now in seventh (and last) release
- ✓ HTTP/1.1
  - Spec published in Nov, 1995
  - Draft IETF standard March 1999
- ✓ HTTP/NG
  - "Next Generation"
  - No current activity

HTTP: What is it?

- ✓ Web's RPC on top of TCP/IP
- ✓ Stateless protocol
- ✓ Separate TCP connection to download each BLOB
  - document with 5 inline images = 6 connections
- ✓ Typed data using RPC
  - Negotiate representation for each connection
  - MIME-like content minus Internet Mail
- ✓ Mime's 7 types:
  - Plain text, audio, video, still images, message, multipart message, application-specific data
- ✓ HTTP 1.1 adds
  - Virtual hosting
  - persistent connections and pipelining
  - Efficient caching
  - Digest Authentication
  - HTTP Extension Framework

The HTTP Request Format



HTTP Request Syntax

```
<method><resource identifier><HTTP version><crlf>
[<Header> : <value>]<crlf>
                .
                .
                .
[<Header> : <value>]<crlf>
     blank line     <crlf>
[Entity body]
```

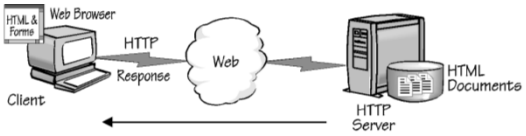
request line
request header fields
entity body

Example

```
GET /path/file.html   HTTP/1.0
Accept: text/html
Accept: audio/x
User-agent: MacWeb
```

request line
request header fields

The HTTP Response Format



HTTP Response Syntax

```
<HTTP Version><result code>[<explanation>]<crlf>
[<Header> : <value>]<crlf>
                .
                .
                .
[<Header> : <value>]<crlf>
     blank line     <crlf>
[Entity body]
```

response header
header fields
entity body

Example

```
HTTP/1.0  200  OK
Server: NCSA/1.3
Mime_version: 1.0
Content_type: text/html
Content_length: 2000

<HTML>
   .
   .
   .
</HTML>
```

response header
header fields
entity body (i.e., HTML document)

For More Information on HTTP

- ✓ See http://www.w3.org/Protocols/

Content Management

- ✓ Manages the content of a Web Site
- ✓ Typically divided into two parts:
  - Content Management Application (CMA)
    - Allows a non-HTML fluent person create, modify, and remove content
    - Tool driven
  - Content Delivery Application (CDA)
    - Uses CMA output to update web site
- ✓ CMS may include:
  - Automation
  - Web Publishing
    - Template/wizard driven
  - Format Management
    - Simplifies import of legacy documents
  - Revision control
  - Indexing, search, retrieval
  - Directed marketing, personalization

Breaking Down the CMS Market

- Segments of the CMS market
  - Document Management Systems (DMS)
    - Optimizes the publishing of documents to any medium including the Web
    - Repository, Metadata, doc history, search, navigation
  - Digital Asset Management (DAM)
    - Similar to DMS but usually for binary files
  - Web content Management
    - Web Content Management (WCM)
      - Adds an additional layer to DMS and DAM
      - Internet and intranet publishing
      - Often integrated with e-commerce applications
  - Learning Content Management (LCM)
    - Produces learning content structured to online learning standards like SCORM and AICC

Market Players

Vendor	Strengths	Caveat
Documentum	DMS and DAM	Weak personalization
Fatwire	WCM, WebLogic integration	Web Scalability
Interwoven	Collaboration, Ent. Content Mgmt	Lots of customization rqd
Percussion	WCM with XML/XSL backbone	Web Scalability
Stellent	Doc to Web conversion	Large scale system
Vignette	Personalization	DMS and library svcs not robust

Open Source Solutions

- ✓ Free or low cost solutions
  - Software low cost but may purchase support, training, or hosting
- ✓ Some examples:
  - Midgard
    - Organizes content into hiearchy ala Yahoo
  - Zope
    - Object oriented App server written in Python
  - Cofax
    - Originally written for Knight-Ridder

For More Info on CMS…

- ✓ See http://cmsInfo.org

The Semantic Web

- The **Semantic Web** is the representation of **data** on the World Wide Web. 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.

"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, Ora Lassila, The Semantic Web, Scientific American, May 2001*

Semantic Web

- The Next Generation Web
  - Coined by Tim Berners-Lee, the originator of the WWW
  - Not a new Web but extension of the current Web that allows people and computers to work together more efficiently
  - Current Web is about *locating* information via *links*
    - No way to add or interpret any semantics of the links
  - Goal of the semantic web is to make the links between information more intelligent
  - Moves from keywords to interpretation of structured information

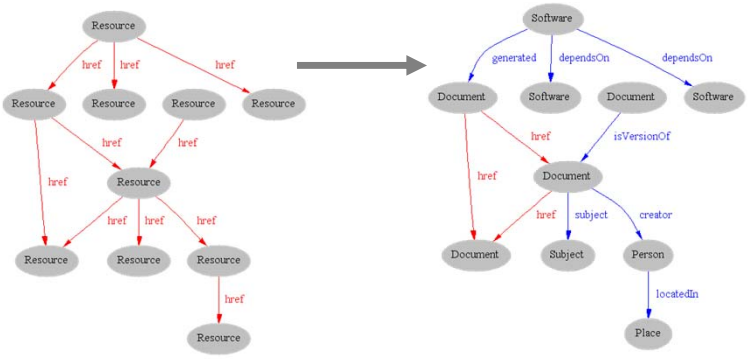
Semantic Web (cont)

- ✓ The intelligent links allow computer interpretation of the links
  - Moves more task management away from computer users
  - Users can have *agents* to perform mundane (and not so mundane) tasks
- ✓ The semantic web will go beyond earlier *artificial intelligence (AI)* systems
  - AI systems have been centralized
    - common concepts
    - Carefully conceived to always produce an exact answer
  - The web is very decentralized…
    - Stifling to have universal definitions
    - Want to have a expressive model for data and rules that can import knowledge from any existing system
      - ❑ May not necessarily produce answers

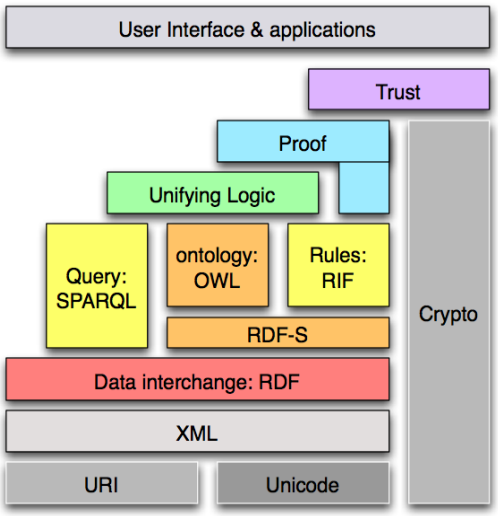
From WWW to the Semantic Web

Web of documents
What do the links mean?

Web of objects
Meaning explicitly represented



The Semantic Web



Technologies and Buzzwords for the Semantic Web

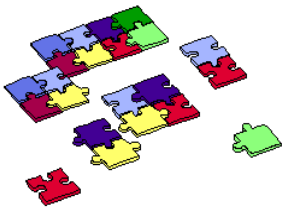
- ✓ XML (eXtensible Markup Language)
  - Allows users to create arbitrary structure for their data
- ✓ RDF (Resource Definition Framework)
  - Allows meaning to be stored as triplets
    - Written using XML
    - Similar to subject, verb, object of sentences
    - *Things* have *properties* with specific *values*
    - Located using URIs
      - Allows new concepts to identified with different URIs
      - Forms a web of information

Technologies and Buzzwords for the Semantic Web (cont)

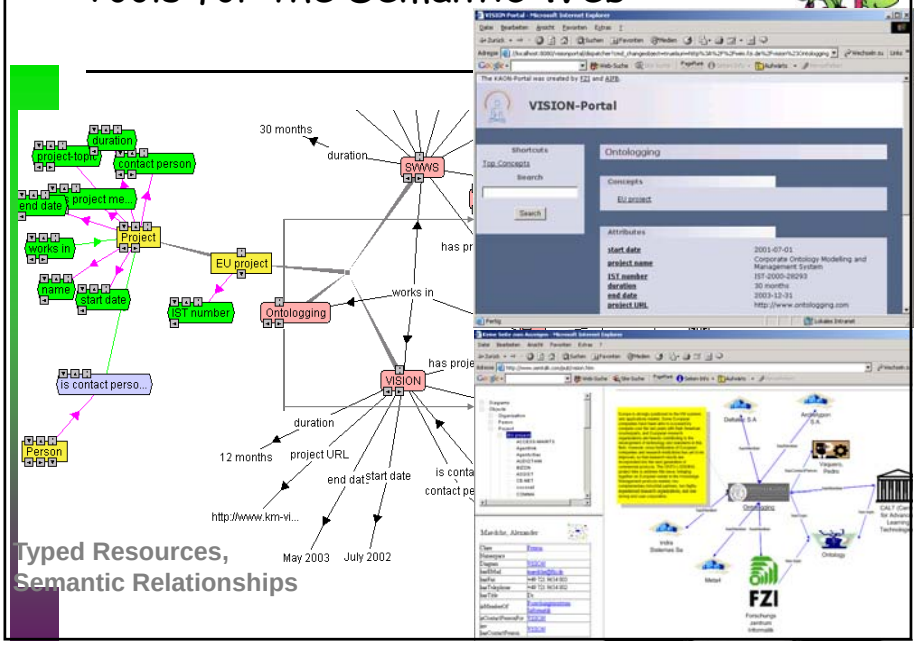
- ✓ Ontologies
  - Third basic component of the Semantic Web
  - Collection of information
  - Defines the relations between terms
  - Typically for the web, a taxonomy and set of inference rules
    - Taxonomy defines classes of objects and relations among them
    - Inference rules allow interpretation of the relations by computers

Technologies and Buzzwords for the Semantic Web (cont)

- ✓ Knowledge Management/Discovery
  - automatically extracting knowledge from data
  - Managing the result
- ✓ Data, Text, and Web Mining techniques provide means for
  - automating the generation o ontologies and metadata
  - contextualizing systems via personalization and user modeling
  - extracting new knowledge, e.g. implicit associations and dependencies, etc.



Tools for the Semantic Web



Typed Resources, Semantic Relationships

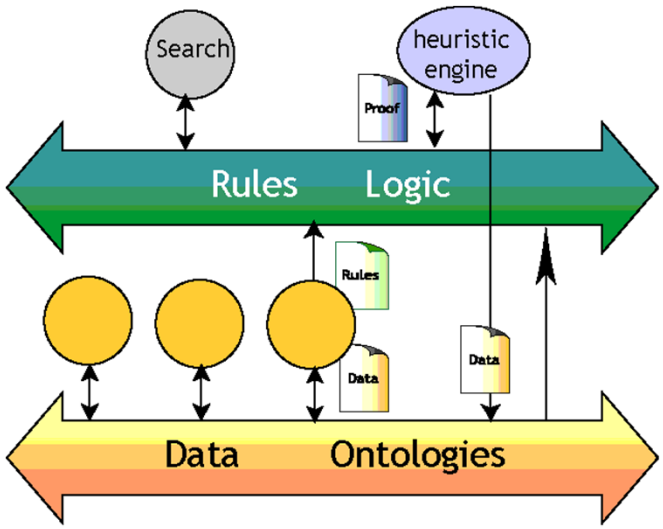
Technologies and Buzzwords for the Semantic Web (cont)

- ✓ Agents
  - Programs that will collect Web content from diverse sources, process the results, then share with other programs
  - Semantic web enables agents
  - Other Web features will be required
    - Semantic exchange languages for proofs (for validating data interpretation)
    - Digital Signatures (data is from trusted sources)
    - Web-based services (Web Services)
    - Service discovery (ala Sun's JINI)

Semantic Services

- ✓ Web Services will play an important role towards using the Web as a basis for high performance work environments.
- ✓ Current Web Service efforts around SOAP, WSDL and UDDI are a starting point to lift the Web to a new level of service.
- ✓ Semantic services will allow for further automatization in a decentralized environment:
  - Discovery (Exploiting P2P services at a semantic level)
  - Mediation (Automatic Translation and Interoperation)
  - Negotiation (Match-making and Comparison)
  - Composition (Dynamic and complex workflows)

The Semantic Web Bus



More On RDF…

- **Framework for describing and exchanging metadata**
  - **Resource** is anything that can have a URI
    - Includes all the Web's pages, as well as individual elements of an XML document
    - Example: http://www.textuality.com/RDF/Why.html
  - **Property** is a Resource that has a name and can be used as a property
    - Example Author or Title
    - Property needs to be a resource so that it can have its own properties
  - **Statement** consists of the combination of a Resource, a Property, and a **value**
    - Known as the 'subject', 'predicate' and 'object' of a Statement
    - Example: "The Author of http://www.textuality.com/RDF/Why.html is Tim Bray."
    - Value can just be a string, for example "Tim Bray" ,or it can be another resource, for example "The Home-Page of http://www.textuality.com/RDF/Why.html is http://www.textuality.com."

- Example Expression in RDF

```
<rdf:Description about='http://www.textuality.com/RDF/Why-
    RDF.html'>
<Author>Tim Bray</Author>
<Home-Page rdf:resource='http://www.textuality.com' />
</rdf:Description>
```

RDF Characteristics

- ✓ Independence
  - any organization can invent Properties
- ✓ Interchange
  - XML allows easy interchange
- ✓ Scalability
  - RDF statements are simple
    - Three part records
    - Easy to look up in large numbers
- ✓ Properties are resources
  - Properties can have their own properties
    - Facilitates searching
- ✓ Values can be resources
- ✓ Statements can be resources

RDF Vocabularies

- ✓ RDF doesn't provide any properties of its own
- ✓ Properties likely to come in groups
  - Example: Author, Title, Date, ...
  - Will come from variety of organizations

Semantic Web: More Information

- ✓ W3C: http://www.w3.org/2001/sw/
- ✓ Variety of Links: http://ai.kaist.ac.kr/~sjcho/semantic-web/
- ✓ RDF: http://www.w3.org/RDF/
- ✓ Collection of articles on Semantic Web: http://www.ercim.org/publication/Ercim_News/enw51/