World-Wide-Web or The Triumph of Anarchy

- Perhaps the most important human technological artifact that evolved more or less ad-hoc
- Limited original vision of the WWW has very little to do with today’s impressive reality
- Web Users have consistently innovated in figuring out new ways of leveraging this powerful medium
- Web architects then try to catch up by extending (read “patching”) the Web infrastructure to support these new uses
- ... this lecture tells their ongoing story!
How it all started...
The Web as a Static Document Repository

- Tim Berners-Lee's original vision for the WWW (circa 1989)
- An easy way to access cross-linked static documents stored in a variety of servers around the world
- Initial specification defined:
  - A language for formatting such documents (HTML)
  - A simple protocol for communicating between browsers and servers (HTTP)

How the (original) Web works

Open Location:
Lecture notes for 15.564: Information Technology I

How the Web works

Open Location:

Looking up host: web.mit.edu ...

Web client
(Netscape, Explorer, etc.)

Internet

Web server

Domain Name Server

What is the IP address of web.mit.edu?

18.30.0.22

Contacting host: web.mit.edu ...

Contacting host: web.mit.edu ...

Web client
(Netscape, Explorer, etc.)

Internet

Web server

Domain Name Server

Here it is

Please send me file /sloan/www/index.html

Here it is

Chrysanthos Dellarocas.
How the Web works

Web client (Netscape, Explorer, etc.)

Internet

Web server

Domain Name Server

WWW is a Client/Server System

- Web Clients
  - Use HTTP protocol to connect to servers
  - Request and display Web pages stored in servers
  - Typical clients: Web browsers
- Web Servers
  - Listen for incoming connections from clients
  - Use HTTP protocol to converse with clients
  - Store and transmit Web pages to clients
Evolution of the WWW

Business Drivers

- **Enable transactions**
- Allow interactivity between browser and server
- Facilitate personalization
- Support multiple browsing devices
- Better organize and retrieve Web content
- Support Business-to-Business applications

The Web as a transaction facilitator

- Business Motivation: Low-cost front-end for allowing customers to connect to corporate computers
  - Customer registration/Address changes
  - Order tracking/Customer support
  - **Online Transactions: eCommerce**!

- Problems of original Web concept
  - Static web pages
  - No interactivity
  - Stateless protocol: no support for multi-step transactions
  - Insecure communications
Example: See FedEx Tracking system at http://www.fedex.com
Web Forms

- Pages which contain fields to be filled by user
- Usually contain a “Submit” button
- When user presses “Submit”, server responds by sending a page containing information specific to the user-supplied parameters
- Examples:
  - Web search tools
  - Order forms in commercial web sites

Web Forms Under the Hood

- Server sends original html page containing input fields
- User types info into fields and presses submit button
- Client establishes connection with handler script at server side (script filename contained in web page)
- Client collects user input into a long string and sends it along with an HTTP command back to server
  - POST customer=John+Doe&cardno=1234567890&expires=6/98&product=123&quantity=5
- Handler script at server reads parameter string and processes it, usually producing a new page as a result
**CGI**

- Common Gateway Interface
- Set of standards for writing handler scripts
- How it works
  - All URLs that refer to a special directory (e.g. /cgi) cause the execution of a corresponding script at the server (for example http://web.mit.edu/cgi/test)
  - Scripts typically translate parameters into SQL statements for a database and translate the query results into an HTML page

---

**Example: FedEx**

- Translate airbill number into query as expected by legacy system
- Package results returned by legacy system into a Web page
Microsoft Active Server Pages (ASP)

• Competing technology to CGI
  - Scripting Language is similar to Visual Basic
• MS Access can automatically convert database tables, queries and forms into ASP pages
• Requires Microsoft web server

Evolution of the WWW

Business Drivers

• Enable transactions
• **Allow interactivity between browser and server**
• Facilitate personalization
• Support multiple browsing devices
• Better organize and retrieve Web content
• Support Business-to-Business applications
The Interactive Web

• Business Motivation:
  – Allow complex interaction between user browser and corporate server
  – Web becomes an extension of the user’s PC
  – Browser becomes a window to a variety of corporate applications

• Problems with Web Forms/CGI/ASP
  – All processing done at server side
  – Rapid user interaction with Web page not possible
  – Need local processing to create highly interactive Web pages

Enter Java Applets

• Programming language to enable interactive Web pages
• Developed by Sun Microsystems
  – originally for programming intelligent microwave ovens!!!
• Java programs are called applets
• Applets are platform-independent
  – They can run equally well on Windows, Macs, Unix, etc.
  – Require special browsers that can support Java though
Java Operation

- Static pages (S)
- Dynamic pages (D)
- Java applets (J)

Client environments

Java Competition: Dynamic HTML

- Extensions to HTML to allow Web pages to be updated without the need to communicate with a Web server
- Browser must be capable of understanding them
- Mutually incompatible versions have been included in Netscape Communicator, Internet Explorer

Chrysanthos Dellarocas.
Evolution of the WWW

Business Drivers

- Enable transactions
- Allow interactivity between browser and server
- **Facilitate personalization**
- Support multiple browsing devices
- Better organize and retrieve Web content
- Support Business-to-Business applications

Personalized Interaction

- Business motivation:
  - Low cost medium for gathering information from customers to allow
    - Personalized service
    - Targeted advertising
- Problems with current model
  - Does not allow easy identification of distinct customers
Cookies

• A method for identifying web users and delivering customized web sites
  - First time user connects to a web site, s/he is asked to fill in personal information form
  - Server packages information into a “cookie” file and sends cookie to browser
  - Browser stores cookie in local file system
  - Each subsequent time browser visits site, it sends cookie back to server
  - Server uses information stored in cookie to identify user and possibly customize the supplied web pages

• Privacy implications?

Cookie applications
Evolution of the WWW

Business Drivers

- Enable transactions
- Allow interactivity between browser and server
- Facilitate personalization
  - Support multiple browsing devices
  - Better organize and retrieve Web content
  - Support Business-to-Business applications

Example: Personalized recommendations at online bookstores.
Multiple Delivery Devices

- Business motivation:
  - Allow users to access web content from a variety of devices
    - PC Browsers
    - PDAs (e.g. Palm Pilots)
    - Mobile Phones
    - Telephones (via voice interface)
    - ...
- Problems of current Web model
  - Each access device has different look-and-feel requirements
  - HTML specifies formatting for PC browsers only

Organize and Index Web Content

- Web is useless unless we can easily locate relevant resources
- Current solution: Search Engines
  - Index the Web by automatically "discovering" web pages and organizing them around keywords found in their text
- Problem:
  - Text keywords are misleading...
  - HTML does not give any clues as to the true meaning of the data
Today’s searches...

Desperately seeking Wendy Cook …

Dirgession:
How Google does it: Pagerank

• Pages are retrieved based on keywords
• Retrieved pages are rank ordered based on a rough measure of “page significance”:
  - How many links point to a page
  - How many links point to the pointing page
  - Etc. for four levels of indirection
• Surprisingly successful in returning relevant hits

### Support for B2B applications

- Original Web was conceived as a communication medium between computers and humans
- Amazing new applications will become possible if computers can automatically read and understand Web pages
  - Electronic purchasing
  - Business intelligence gathering
  - ….

### Problem:
- HTML pages are unstructured
- HTML only provides information about presentation, not meaning

### What is the underlying issue?

- When storing documents on the web, specify not only their appearance, but also their semantics (i.e. their meaning!)
Enter: The Semantic Web

- The "Next Generation Web" with well-established infrastructure for expressing information in a
  - Precise,
  - Human-readable, and
  - Machine-interpretable form.
- Enable syntactic and semantic interoperability among independently-developed Web applications, allowing them to efficiently perform sophisticated tasks for humans.
- Enable Web resources to be accessible by their semantics rather than by keywords and syntactic forms.
- Enable inferencing:
  - Chris is an associate professor at MIT.
  - Associate professors are permanent employees.
  - Chris is a permanent employee of MIT.

The Origins of the Semantic Web

- The information retrieval crisis beginning in the late 1990s led to a widespread interest in what has come to be called metadata.
- What is metadata?
  - It's just data.
  - But it's data about other data
  - Data intended for machine consumption
- What could metadata do for us?
  - Give search engines something to work with (relational triples) that is designed for their needs.
  - Give us all a place to record what a document, or any other resource, is for or about.
First Requirements for Metadata

• What would we need to make this work?
  – A standard syntax, so metadata can be recognised as such;
  – One or more standard vocabularies, so search engines, producers and consumers all speak the same language;
  – Lots of resources with metadata attached;

HTML = Hypertext Markup Language

• Uses tags to specify the formatting and display properties of data
  – Sizing of Fonts
  – Indentation
  – Etc.
• Says nothing about what data means...

<h1>March 25, 1998 08:00</h1>

<blockquote>
<h2><i>Seattle, WA</i><br/>
</h2>
</blockquote>

<h2>Partly Cloudy Skies</h2>

<h2>Temperature: 46 degrees</h2>
XML = eXtensible Markup Language

- Represents data as a list of hierarchically structured fields
- Syntax similar to HTML
- Uses tags to delimit the <beginning> and </end> of a field
- Can be easily extended with more tags to represent additional kinds of data

```
<weather-report>
  <date>March 25, 1998</date>
  <time>08:00</time>
  <area>
    <city>Seattle</city>
    <state>WA</state>
  </area>
  <measurements>
    <skies>partly cloudy</skies>
    <temperature>46</temperature>
  </measurements>
</weather-report>
```

XML is just a syntax...

- ... for describing the meaning of data stored on a Semantic Web page
- In order for XML to be useful, organizations must agree on common **ontologies**
  - Concepts of interest to their domain
    - E.g. product, quantity, price, weight, size, delivery date
  - Relationships among concepts
    - Product-has-price, Order-has-delivery date
  - Vocabularies of XML tags to represent the above

- This is an organizational not a technical problem!
Example: RosettaNet

An industry consortium whose objective is to develop XML-based B2B supply chain management infrastructure in Information Technology, Electronic Components and Semiconductor Manufacturing.

For company description, see http://www.rosettanet.org

Rosettanet Objectives

For company objectives, see http://www.rosettanet.org
Industry support for XML

- Organizations developing standard vocabularies
  - commerce.net -- e-commerce processes
  - rosettanet.org -- supply chain processes
- Vendors incorporating XML support in their products
  - Microsoft -- XML support in IE 5.0 and Office 2000
  - Oracle -- XML support in databases
  - IBM -- leading in XML parsers and viewers
- XML-based business2business networks
  - Ariba procurement network
  - New York times syndication network
  - IFX (Charles Schwab and others)

Semantic Web Technological Layers

- **XML** Customized tags, like:
  `<dog>Nena</dog>`
- **RDF** Relations, in triples, like:
  `(Nena) (is_dog_of) (Chris)`
- **Ontologies** Hierarchies of concepts, like
  `animal --> mammal --> dog`
- **Inference rules** Like:
  If (person) (owns) (dog), then (person) (cares_for) (dog)

= **Semantic Web!**
Semantic Web Applications: Better Searches


Semantic web applications:

Intelligent Agents

Will this ever happen?