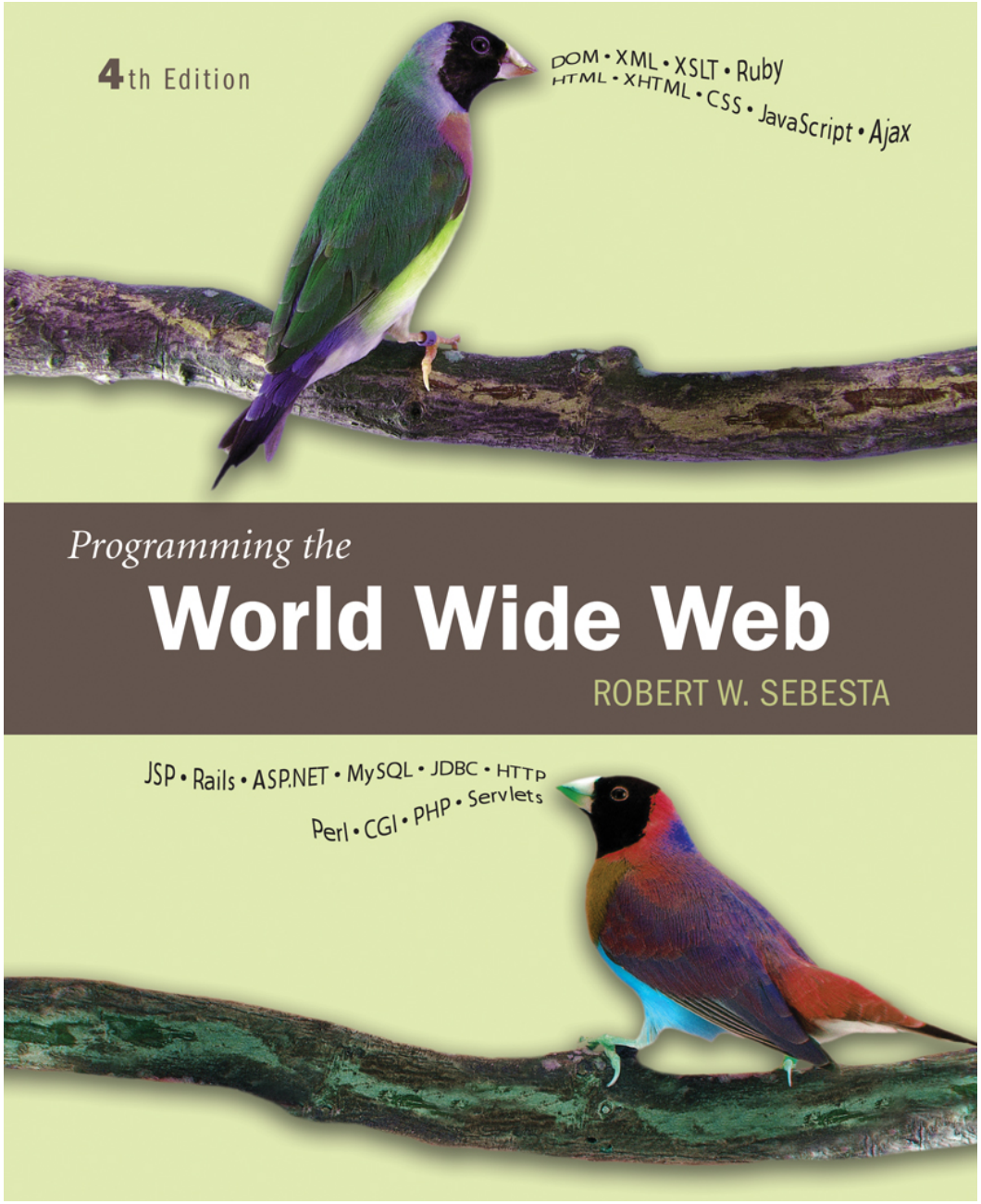


Chapter 1

Fundamentals

Introduction to Web Programming

Objectives

- To understand the technology and protocols underlying the World Wide Web(WWW)

- To become familiar with common tools and techniques for developing Web-based applications, both client-side and server-side

- To develop a working knowledge of HTML, XHTML, JavaScript, Java, Perl and PHP as languages for developing Web applications

Learning Outcomes

- ✔ Demonstrate an understanding of the concepts, terms, and technology behind the WWW
- ✔ Describe how the WWW works
- ✔ List several Web servers
- ✔ Identify different Web browsers
- ✔ Define what is a Web page
- ✔ Differentiate a home page from a Web site
- ✔ Understand how a Web page works
- ✔ Static and Dynamic pages
- ✔ HTTP versions, methods, return status codes
- ✔ Overview of Web programmer's tools

Internet History

- Origins
  1. ARPAnet - late 1960s and early 1970s
     - Network reliability
     - For ARPA-funded research organizations

  2. BITnet, CSnet - late 1970s & early 1980s
     - email and file transfer for other institutions

  3. NSFnet - 1986
     - Originally for non-DoD funded places
     - Initially connected five supercomputer centers
     - By 1990, it had replaced ARPAnet for non-military uses
     - Soon became the network for all (by the early 1990s)

  – NSFnet eventually became known as the Internet

What the Internet is:

- A world-wide network of computer networks
- At the lowest level, since 1982, all connections use TCP/IP
- TCP/IP hides the differences among devices connected to the Internet

Internet Protocols

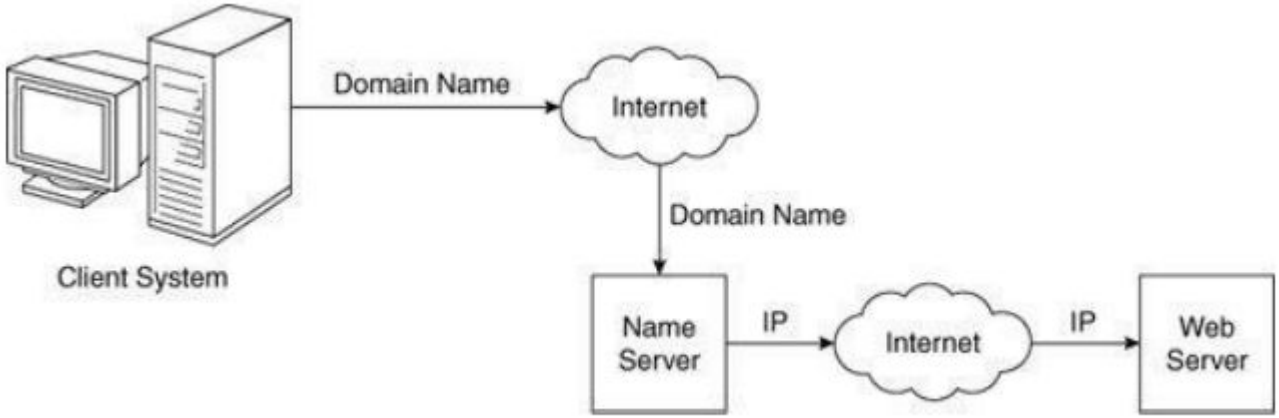
- Internet Protocol (IP) Addresses
  - Every node has a unique numeric address
  - The Internet Protocol (IP) address of a machine connected to the Internet is a unique 32-bit number
  - Form: 32-bit binary number
    - New standard, IPv6, has 128 bits (1998)
    - www.dsce.edu.in
- Organizations are assigned groups of IPs for their computers.

- Problem: By the mid-1980s, several different protocols had been invented and were being used on the Internet, all with different user interfaces (Telnet, FTP, mailto).

Internet Protocols

- Domain names
  - Form: host-name.domain-names
  - First domain is the smallest; last is the largest.
  - Last domain specifies the type of organization.
  - Fully qualified domain name - the host name and all of the domain names.
  - DNS servers - convert fully qualified domain names to IPs.

Eg: movies.comedy.marxbros.com



Web ≠ Internet

Internet

a physical network connecting millions of computers using the protocols for sharing/transmitting information (TCP/IP)

- in reality, the Internet is a network of networks
- basically, just a computer network spanning most of the world

World Wide Web

a collection of interlinked multimedia documents that are stored on the Internet and accessed using a common protocol called Hyper Text Transfer Protocol (HTTP)

an immense source of data, enables to access the internet

a collection of software and protocols, installed on most of the computers on the internet

Key distinction: Internet is hardware; Web is software

History of the Internet

The idea of a long-distance computer network traces back to early 60's

- Licklider at M.I.T.
- National Physics Laboratory in U.K.

In particular, the Department of Defense(DoD) was interested in the development of distributed, decentralized networks

- survivability (i.e., network still functions despite a local attack)
- fault-tolerance (i.e., network still functions despite local failure)

In 1969, Advanced Research Project Agency funded the ARPANET

- connected computers at UCLA, UCSB, SRI, and Utah
- allowed researchers to share data, communicate thru simple text based e-mail

In late 1970s and early 1980s

- BITNET (Because It's Time Network) developed at City University of New York
- CSNET (Computer Science Network) connected Purdue Univ, Univ of Wisconsin etc
- Neither BITNET nor CSNET became a dominant national network

In 80's, U.S. government took a major role in Internet development

- created NSFnet(National Science Foundation) for academic research in 1986
- ARPANET was retained for military & government computers

By 90's, Internet connected virtually all colleges & universities

- businesses and individuals also connected as computing costs fell
- ~1,000,000 computers by 1992 were connected around the world

In 1995, control of the Internet was transferred to a non-profit organization

- Internet Society: Internet Engineering Task Force
  Internet Architecture Board
  Internet Assigned Number Authority
  World-Wide-Web Consortium
  . . .

Internet has exhibited exponential growth, doubling in size every 1-2 years
(stats from *Internet Software Consortium)*

1,463,632,361 million Internet users in 2008 (approx. 22% of the world's population)
(http://www.internetworldstats.com/top20.htm)
(June 30, 2008)
in India, 81,000,000(Users) / 1,147,995,898 (Population)
81%

United Kingdom has 41.8 million users (approx. 69% of the population)

Year	Computers on the Internet
2006	439,286,364
2004	285,139,107
2002	162,128,493
2000	93,047,785
1998	36,739,000
1996	12,881,000
1994	3,212,000
1992	992,000
1990	313,000
1988	56,000
1986	5,089
1984	1,024
1982	235

The World-Wide Web

- A possible solution to the proliferation of different protocols being used on the Internet

- Origins
  - Tim Berners-Lee at CERN proposed the Web in 1989
    - Purpose: to allow scientists to have access to many databases of scientific work through their own computers
  - Document form: hypertext
  - Pages? Documents? Resources?
    - We'll call them documents
  - Hypermedia – more than just text – images, sound, etc.

- Web or Internet?

  The Internet is a collection of computers and other devices connected by equipment that allows them to communicate with each other.

  The Web is a collection of protocols that has been installed on most of the computers on the internet.

  Web uses one of the protocols, http, that runs on the Internet--there are several others (telnet, mailto, etc.)

What is the World Wide Web?

- Information resource consisting of web pages that organize and present vast amount of information (mostly text embedded with images, audio, video, or animation), and other resources (databases, interactive multimedia, virtual environments, etc.)
- A hypertext based system for providing, organizing and accessing information that allows users to jump from one information space to another
- A way to access and provide information in various media via the Internet
- Comprises servers and client computers on the Internet that communicate using the hypertext transfer protocol (http)
- Body of information available on the Web
- An easy way to access cross-linked static/dynamic documents stored in a variety of servers around the world
  - A language for formatting such documents (HTML)
  - A simple protocol for communicating between browsers and servers (HTTP)

History of the Web

The idea of hypertext (cross-linked and inter-linked documents) traces back to Vannevar Bush in the 1940's

- online hypertext systems began to be developed in 1960's
  e.g., Andy van Dam's FRESS, Doug Englebert's NLS
- in 1987, Apple introduced HyperCard

In 1989, Tim Berners-Lee at the European Particle Physics Laboratory (CERN) designed a hypertext system for linking documents over the Internet

- designed a language for specifying document content
  - which evolved into HyperText Markup Language (HTML)
- designed a protocol for downloading documents and interpreting the content
  - which evolved into HyperText Transfer Protocol (HTTP)
- implemented the first browser -- text-based, no embedded media

the Web was born!

The Web was an obscure, European research tool until 1993

In 1993, Marc Andreessen (at the National Center for Supercomputing Applications) developed Mosaic, the first graphical Web browser

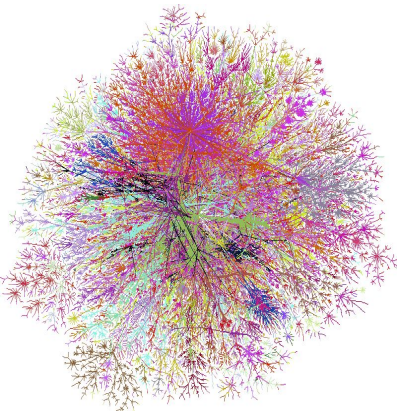
- the intuitive, clickable interface made hypertext accessible to the masses
- made the integration of multimedia (images, video, sound, …) much easier

- Andreessen left NCSA to found Netscape in 1994
  - cheap/free browser popularized the Web (75% market share in 1996)
  - in 1995, Microsoft came out with Internet Explorer
  - Netscape bought by AOL in 1999 for $10 billion in stock

Today, the Web is the most visible aspect of the Internet

Web Growth

Stats from Netcraft Web Server Survey.



Different colors represent the proliferation of different IP addresses

Firefox

IE, Opera

Netscape

Mosaic

Year	Computers on the Internet	Web Servers on the Internet
2004	285,139,107	56,923,737
2002	162,128,493	33,082,657
2000	93,047,785	18,169,498
1998	36,739,000	4,279,000
1996	12,881,000	300,000
1994	3,212,000	3,000
1992	992,000	50

Recent estimates suggest 135 million Web servers, with 4-5 B Web pages!
Note: Growth of web is around 1 million pages/day. Roughly 10 times the rate of population growth in India.

What makes the Web work?

The Web relies on these mechanisms:

- Hypertext - provides easy navigation among documents and resources
- Protocols - set of standards used to access resources via the Web
- Universal Resource Locator (URL) - uniform naming scheme for Internet resources
- Client and Server computers - Web access is based on client/server technology

Hypertext

- Presents and relates information as hyperlinked documents that point to other documents or resources
- Hyperlink is usually embedded in the text, on a highlighted word or phrase, or on a symbol, an icon, or other graphic elements
- Web pages are hypertext documents on the Internet mostly created using HTML

HyperText Markup Language (HTML)

- The publishing language of the World Wide Web; the standard used to create web pages.
- Defines the structure of information by using a variety of tags and attributes, which is designed to display text and other information on a screen and provide hyperlinks to other Web documents.
- Defines a standard set of special textual indicators(markups) specifying how a Web pages, words and images should be displayed by the web browser

Protocols

- Standard set of rules that governs how computers communicate with each other, i.e. SMTP, FTP, HTTP allowing access to huge collection of information and services

Uniform Resource Locator (URL)

- Uniform naming scheme that specifies unique addresses for web servers, documents, and other resources, no matter what its access protocol
- Web resources that are accessible though the Internet are identified by URLs
- A URL looks like this: http://www.dsce.in/ise.html
- It is composed of three parts. The start, http://, indicates that this URL uses the HTTP protocol. The next part, www.dsce.in, names the server on which this page can be found. The end of the URL, /ise.html, names a specific file on this server
- Web documents and resources are located and linked through their URLs

Anatomy of a URL

Note: Not all URLs will have the directory and filename

Client and Server computers

- The web operates in Client/Server Configuration
- In HTTP communication, the server is the computer on which the web-page is stored. The client is the computer, which asks the server for a page, so that it can display it. asking for a page is called an 'HTTP request'.
- Servers are computers that host web documents and provide information through a web server program
- Client computers access web documents using an application program called web browsers
- Client initiates the communication, i.e., a request for information on the server, Server sends information back to the client

What is a Web Browser?

- Client-side software that is responsible for displaying page views and making HTTP requests to a web server.

- Application software that is used to locate and issue a request for the page on the web server that hosts the document.
- It also interpret the page sent back by the web server and display it on the monitor of the client computer.

- Computer program that lets you view and explore information on the World Wide Web. i.e., Allow the users to browse the resources available on servers.

Browsers are clients - always initiates the communication, servers react (although sometimes servers require responses).

- Most requests are for existing documents, using HyperText Transfer Protocol (HTTP).
- But some requests are for program execution, with the output being returned as a document.

Mosaic - NCSA (Univ. of Illinois), in early 1993

First to use a GUI, led to explosion of Web use

Initially for X-Windows, under UNIX, but was ported to other platforms by late 1993.

Web browsers : IE, Firefox, Mozilla, Netscape Navigator, Opera.

More popular: IE and Firefox

Web Browsers : Examples

Microsoft Internet Explorer – browser integrated with the Windows operating system. Mac versions are available.

Netscape Navigator - available for Windows, Mac, and Unix platforms.

Opera – one of the alternatives to the two most popular browser mentioned above.

Mozilla – open source web browser software.

Firefox – available in versions for several different computing platforms such as Windows, Linux, Mac OS.

What is a Web Server?

- Provides access to the web resources.
- Server-side software responsible for handling incoming HTTP requests.
- Computer running application software that listens and responds to a client computer's request made through a web browser.
- Machine that hosts web pages and other web documents.
- Most commonly used Web Servers are: Apache, IIS.
- More than 400 million web hosts in operation, more than 60% of which were Apache, around 30% were IIS.

- Provide responses to browser requests, either existing documents or dynamically built documents.

- Browser-server connection is now maintained through more than one request-response cycle.

- All communications between browsers and servers use Hypertext Transfer Protocol (HTTP).

Web Servers : Examples

- Apache - most popular open source server software on the Web.
- iServer – application / web server written entirely in Java.
- Microsoft Internet Information Server - IIS is fully integrated into the Windows NT / 2000 server package.
- Macromedia ColdFusion – application / web server focuses on serving dynamic pages supporting other Macromedia products like Flash and Ultradev.
- IBM Web Sphere Studio – combination of content creation software with web application.
- Apple Webobject - application / web server for Mac

Web Server Operation

- Web browsers initiate communications with servers by sending them URLs.

- A URL can specify one of the two different things:
  - the address of a data file stored on a server that is to be sent to the client.
  - a program stored on the server that the client wants executed.

- All communications between a Web client and a Web Server use the standard web protocol, Hypertext Transfer Protocol(HTTP).

- The primary task of a Web Server is to monitor a communication port on its host machine,

  accept HTTP commands through that port,

  and perform the operations specified by the commands.

- HTTP commands include a URL, which includes the specification of a host machine.

- When the URL is received it is translated to either a filename or a program name.

- All current web servers have a common ancestry.

  The first two servers, developed at CERN in Europe and NCSA at the University of Illinois.

General Server characteristics

- File structure of a web server has 2 separate directories.

- Root of one of these - Document root .
- File hierarchy that grows from the document stores the web documents to which the server has direct access and normally serves the clients. Root of the other directory – Server root.

- Files stored directly in the document root are those available to the clients through top-level URLs.

- Server maps the clients requested URLs to the document root, whose location is not known to clients.

- Many servers allow part of the servable document collection to be stored outside the directory at the document root.
- The secondary areas from which documents can be served are called virtual document trees.

- Servers are now able to support more than one site on a computer, potentially reducing the cost of each site and making their maintenance more convenient. Such secondary hosts are called virtual hosts.

- Some servers can serve documents that are in the document root of other machines on the web, they are called proxy servers.

- In addition to HTTP protocol, web also supports FTP, Gopher, News and mail.

- Apache web server:

  Is an excellent server because it is both fast and reliable.

  It is open-source software.

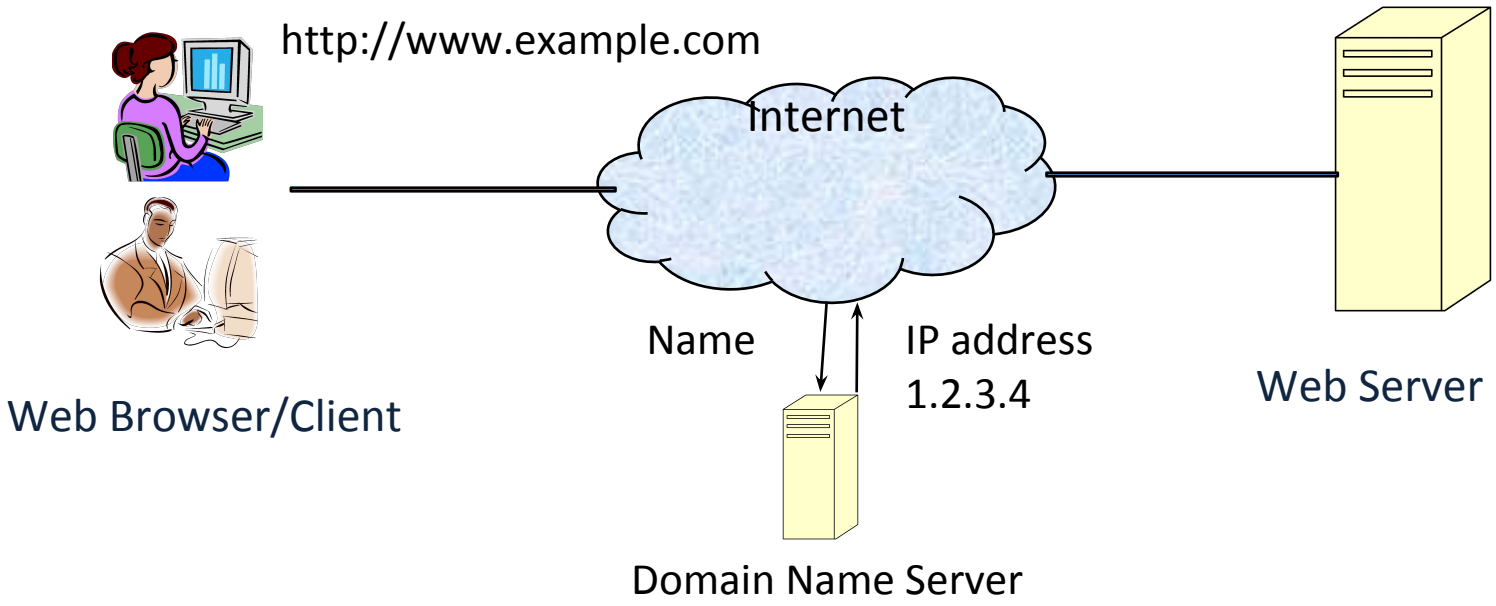
  It is capable of providing a long list of services beyond the basic process of serving documents to clients.

- When Apache begins execution, it reads its configuration information from a file and sets its parameters to operate accordingly.
- http.conf actually stored the directives that controls an Apache server's behavior.

- Microsoft Internet Information Server (IIS) is supplied as part of windows.

What happens when we click on a link? OR enter a URL into the location box OR request a web page

→ How the Web works



-Render HTML images
-Execute JavaScripts
-Execute Java Applets
-Send data to server using CGI

-Serve up Content
-HTML, Images, Documents etc.,
-Process data received from client

How the Web Works

When we click a link or type a URL into location box,

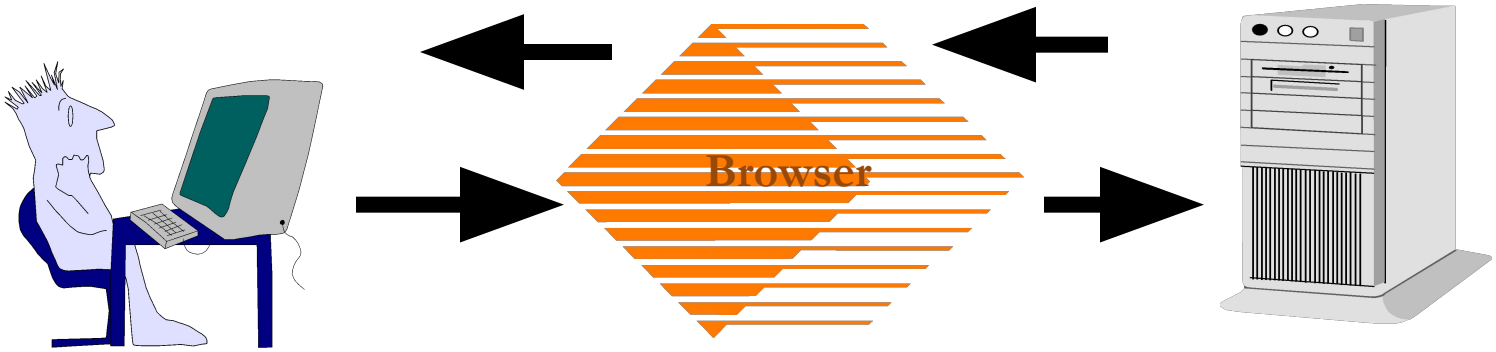
ex: http://www.example.com,

1. Browser makes a socket connection to the server www.example.com
2. This name www.example.com maps to an IP address using DNS, ex:1.2.3.4(dotted quad)
3. Browser connects to www.example.com using port 80, that the server OS opens for such http requests. Note: ftp – 21, telnet – 23, smtp – 25
4. Server serves up or deliver information to the client like HTML file, images, java applets, documents, PDFs etc.,
5. The content that the server delivers can be generated by the server in one of the several ways: static, dynamic, or embedded
6. Client receives a stream of text, images, documents etc., from server and render or appropriately display them. Also, it may execute java applet or JavaScripts.

How the Web works - Summary

5. User receives file displayed by the browser

4. Server sends requested files to browser to be interpreted

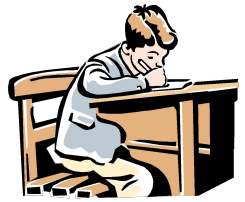


1. User sends request

2. Browser interprets user's selection and makes request from appropriate server

3. Server accepts and processes request from browser

Exercise



Visit the following to know more about how the World Wide Web works.

- Client/Server, the Internet, and WWW
  http://www.robelle.com/www-paper/paper.html
- How Web servers and the Internet Work
  http://www.howstuffworks.com/web-server.htm
- The Web At-a-glance
  http://www.learnthenet.com/english/web/000www.htm

Explore these pages that contain links to several web servers and browsers:

- World Wide Web Server Software
  http://www.w3.org/Servers.html
- Browsers
  http://www.webreference.com/internet/software/browsers/

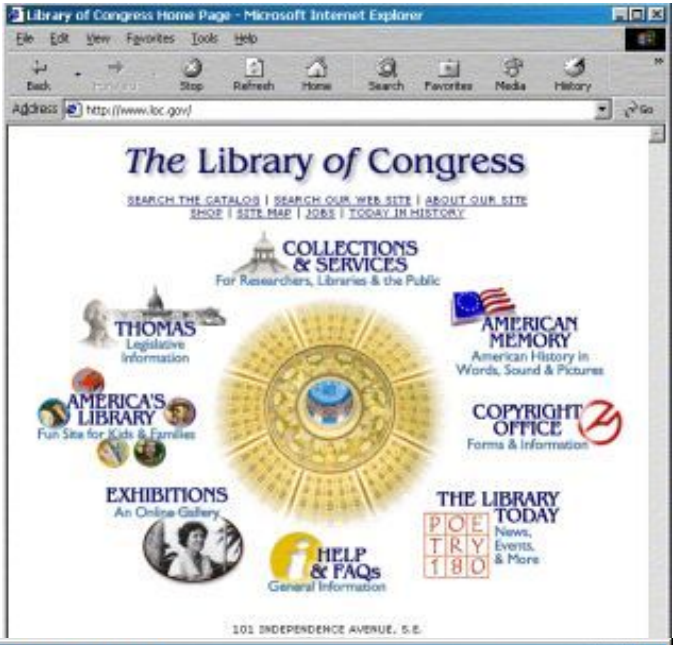
What is a Web page?

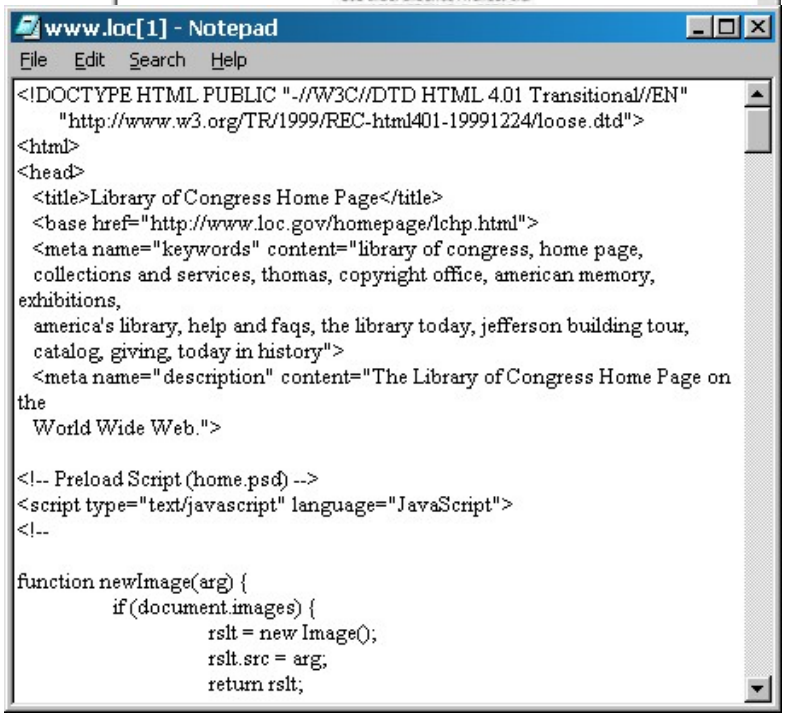
- Set of data consists of one or several web resources, that can be identified by an URI(Universal/Uniform Resource Identifier).
- Electronic document that typically contains several types of information accessible via the World Wide Web .
- set of information created, and organized, using HTML and/or other web page authoring and development tools .
- Interpreted and displayed on the screen according to the instructions of the web page authoring tool

Sample Web Page

Sample web page and its source.

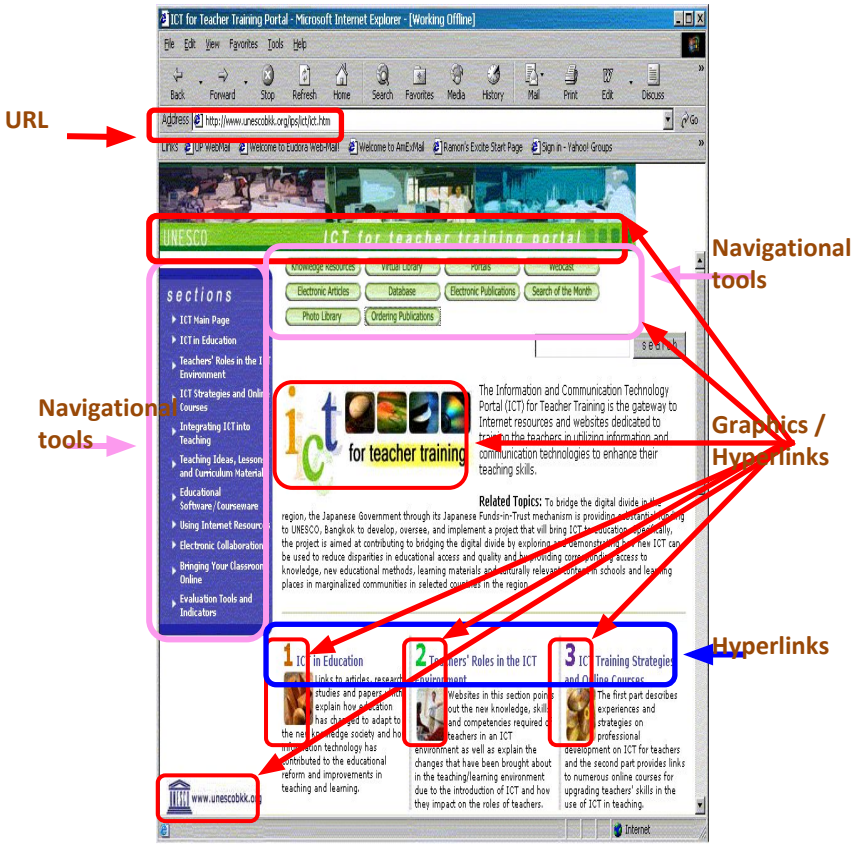
- The source contains the instructions that define the contents, layout, and structure of a web page.
- The instructions are written in HTML or another web authoring tool used in creating the page.
- The browser uses these instructions to interpret and display the web page on the screen.





How web page works

- The user requests a web page by entering its URL on the address location bar of a web browser.
- The browser transmits the request to a web server through http.
- The web server processes the request, locates and sends back the requested web document using http.
- The web browser interprets the file sent by the server and displays it on the monitor.
- The same process happens when the user selects any of the hyperlinks or navigational tools on the page.



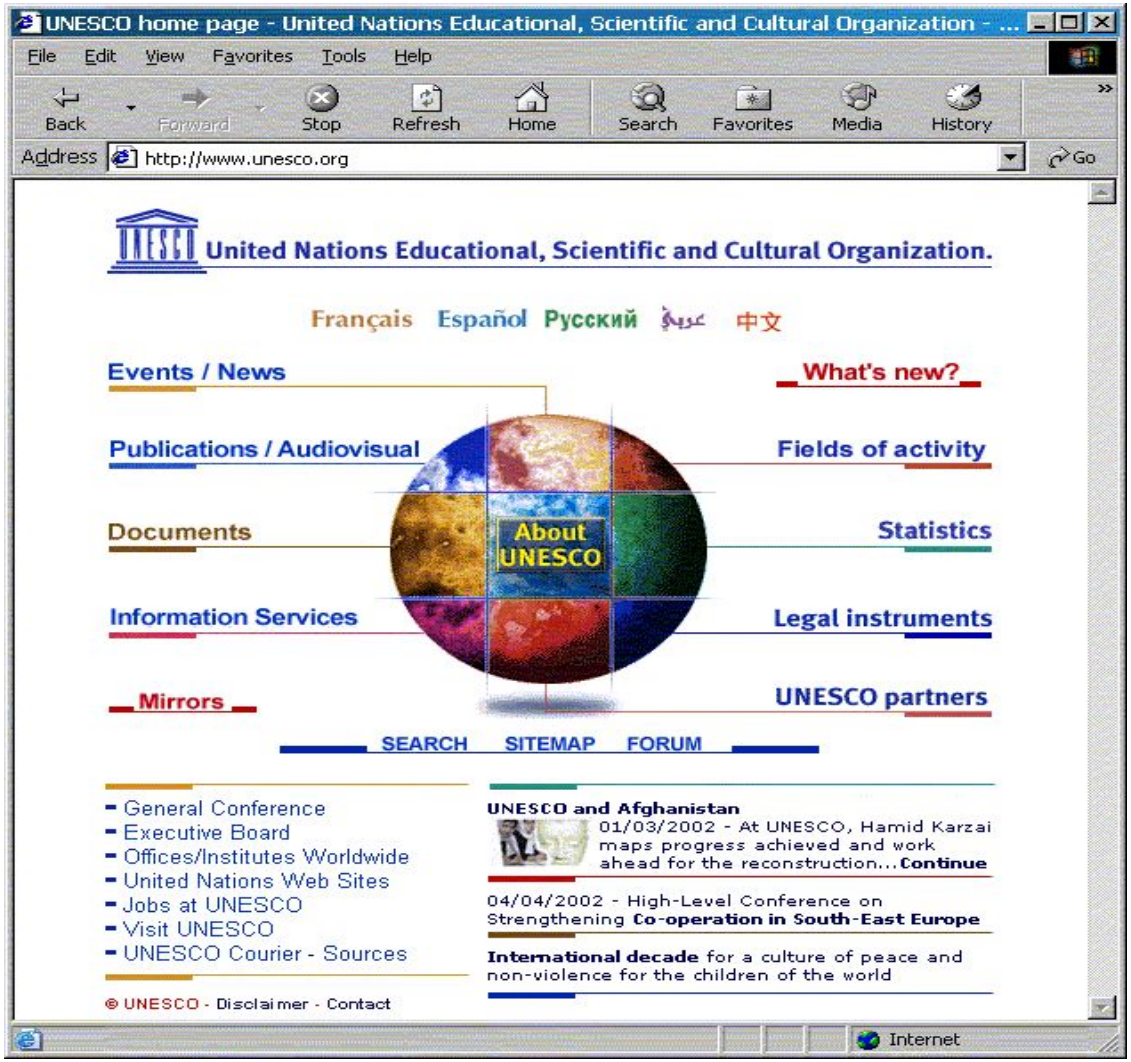
What is a Web Site?

- a collection of related web pages of a certain individual, group, or organization, connected through a system of hyperlinks, hosted in a particular domain.
- can be a single web page that contains links to related information located on several web sites.

What is a home page?

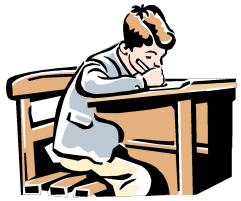
- the main page of a web site that typically serves as an index or table of contents to other web pages
- usually the first web page or the welcome page the users see when they visit a web site

Web page? Web site? Home page?



This web page is the home page of the UNESCO web site

Exercise



Read these articles:

- Dreamlink: What is a web page?

  http://www.dreamink.com/beginners/b2.html
- How web pages work

  http://www.howstuffworks.com/web-page.htm
- How Web Sites Work

  http://www.workz.com/content/629.asp

Multipurpose Internet Mail Extensions (MIME)

- Originally developed for email
- Used to specify to the browser the form of a file returned by the server (attached by the server to the beginning of the document)
- Type specifications
  - Form:

    type/subtype
  - Examples: text/plain, text/html, image/gif, image/jpeg

- Server gets type from the requested file name's suffix (.html implies text/html).

- Browser gets the type explicitly from the server.

- Experimental types
- Subtype begins with x-

  e.g., video/x-msvideo

- Experimental types require the server to send a helper application or plug-in so the browser can deal with the file.

HTTP (Hypertext Transfer Protocol)

- generic, stateless protocol
  - Does not provide storing of information between requests.
  - No indication of any relationship between two different requests.
    - cookies, small data structures that a web server requests the HTTP client to store on the local machine, are used to maintain state information.
- governs the transfer of files across a network.
- developed at CERN (Central European Research Network), they also came up with the name WWW, later W3C.
- supports access to SMTP,FTP and other protocols.
- was designed to support hypertext.
- exchanged information, can be static or dynamic.
- based on client/server model typically using TCP/IP sockets.
- Consists of two phases: Request phase and Response phase.

HTTP (Hypertext Transfer Protocol)

- Request Phase
  - General form of an HTTP request
    - HTTP Method Domain part of URL HTTP version
      Followed by header fields, blank line, and message body
    - Ex: GET /content/chapter1.html HTTP/1.1
  - HTTP Request methods

Method	Description
GET	Returns the content of the specified document i.e., Retrieve document or document produced by a program
HEAD	Returns the header information for the specified document
POST	Executes the specified document, using the enclosed data i.e., Append or attach information
PUT	Replaces the specified document with the enclosed data i.e., Store information
DELETE	Deletes the specified document i.e., Delete the resource indicated in the request

HTTP (Hypertext Transfer Protocol)

- Response Phase
  - General form of an HTTP response : Status Line follwed by response header fields, blank line, and response body
    - » HTTP version 3-digit status code short textual explanation for the status code
    - Ex: HTTP/1.1 200 OK
  - HTTP Status Codes

Code	Description
1XX	Informational, No 1xx status codes are defined, and they are reserved for experimental purposes only
2XX	Success, Means that the request was processed successfully 200 OK Means that the server did whatever the client wanted it to, and all is well
3XX	Redirection, Means that the resource is somewhere else and that the client should try again at a new address, 301 Moved permanently, 302 Moved temporarily
4XX	Client error, Means that the client screwed up somehow, usually by asking for something it should not have asked for. 400: Bad request , 401: Unauthorized, 404: Not found
5XX	Server error, means that the server screwed up or that it couldn't do as the client requested, 500: Internal server error , 503: Service unavailable

HTTP (HyperText Transfer Protocol) - Versions

- Three versions of HTTP.
  - The first one was HTTP/0.9, which was truly primitive and never really specified in any standard.
  - This was corrected by HTTP/1.0, which was issued as a standard in RFC 1945. this is in common use today
  - RFC 2068 describes HTTP/1.1, which extends and improves HTTP/1.0 in a number of areas. Very few browsers support it.

- The major differences are a some extensions in HTTP/1.1 for authoring documents online via HTTP and a feature that lets clients request that the connection be kept open after a request so that it does not have to be reestablished for the next request. This can save some waiting and server load if several requests have to be issued quickly.

Client-side programming

Can download program with Web page, execute on client machine

- simple, generic, but insecure

JavaScript

- a scripting language for Web pages, developed by Netscape in 1995
- uses a C++/Java-like syntax, so familiar to programmers, but simpler
- good for adding dynamic features to Web page, controlling forms and GUI
- requires users to have this technology enabled on their browsers
- see http://www.w3schools.com/js/

Java applets

- can define small, special-purpose programs in Java called applets
- provides full expressive power of Java (but more overhead)
- good for more complex tasks or data heavy tasks, such as graphics
- see http://java.sun.com/applets/

Server-side programming

Can store and execute program on Web server, link from Web page

- more complex, requires server privileges, but secure

CGI programming

- programs can be written to conform to the *Common Gateway Interface*
- when a Web page submits, data from the page is sent as input to the CGI program
- CGI program executes on server, sends its results back to browser as a Web page
- good if computation is large/complex or requires access to private data
- we will discuss CGI programming using Perl, but other languages possible as well (such as PHP, Python, Ruby, etc.)

Active Server Pages, Java Servlets, PHP, Server Side Includes

- vendor-specific alternatives to CGI
- provide many of the same capabilities but using HTML-like tags
- some of these technologies might require functionality to be enabled in the client's browser (e.g. Ajax generally requires the use of Javascript)

Server-side programming v/s Client-side web programming

Server-side

- Server-side scripts or programs are simply programs that are run on the web server in response to requests from the client.
- Technologies such as CGI, Active Server Pages, Java Servlets, PHP run on Server.
- Best, if the program needs a lot of data and infrequent interactions with the server.

Client-side

- Client-side scripts or programs are simply programs that are run on the web browser, downloaded from server.
- Technologies such as JavaScript, VBScript and Java applets all run in the client.
- Best, if applications that use less data and more interactions.

Static v/s Dynamic pages

Most Web pages are *static*

- contents (text/links/images) are the same each time it is accessed

  e.g., online documents, most homepages

Hypertext Markup Language (HTML) is used to specify text/image format.

As the Web moves towards online services and e-commerce, Web pages must also provide *dynamic* content.

- pages must be fluid, changeable (e.g., rotating banners).
- must be able to react to the user's actions, request and process info, tailor services.

e.g., amazon.com

Exercise

Pick some of your favorite Web sites and try to identify

- static components?
- dynamic components?
  - client-side? JavaScript? Java applet?
  - server-side? CGI? ASP?

Ex: www.amazon.co.uk

Web Programmer's Toolbox

Common Tools used in Web Programming:

- XHTML, a markup language.
- XML, a meta-markup language.
- JavaScript, Java, Perl, PHP and Ruby, programming languages.
- Ajax, a web technology that uses JavaScript and XML.
- Rails, a development framework for web based database access systems.

Web programs and scripts are divided into two categories:

- Client side (XHTML, XML, JavaScript, Java applets).
- Server side (Perl, PHP, Ruby, Servlets).