1. Internet Basics for Web Engineering

a. Introduction to the Internet and the Web

b. HTTP and HTTPS

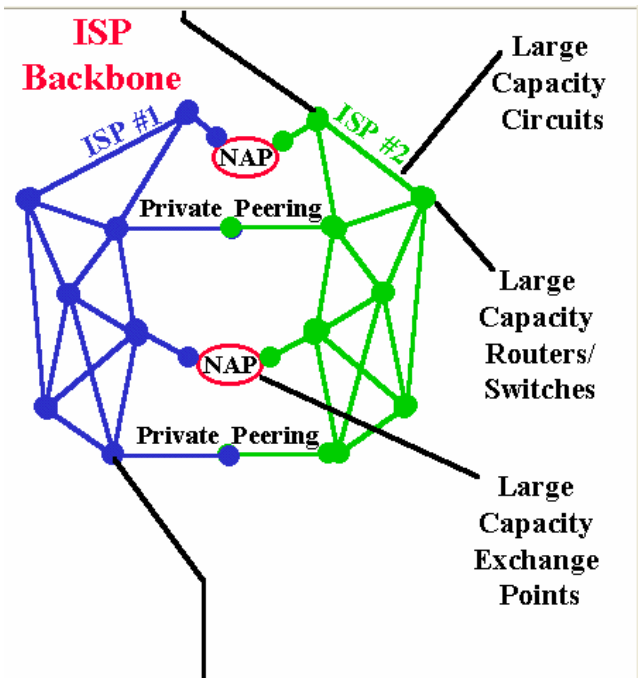
c. Internet Search

1.1 Introduction to the Internet and the Web

- The Internet versus the World Wide Web
  - Definition
  - Brief history
  - Structure/architecture
- Standards
- Web Browsers and Compatibility Issues

What is the Internet?

- A collection of interconnected *computer networks*, linked by copper wires, fiber-optic cables, wireless connections, etc.
  - The transport vehicle for the information stored in files or documents on another computer.
- These networks rely on NAPs, backbones and routers to talk to each other.



What is the Internet?

- Network access points (NAPs): data communications facilities that provide access to higher-speed links (typically intercontinental in extent)
- Backbones: the high-speed, main trunk connections that carry Internet traffic
  - The ISP backbone interconnects the ISP's POPs, AND interconnects the ISP to Other ISP's and online content.
  - The backbones meet at NAPs
- Routers: computer networking devices that forward data packets across a network toward their destinations
- WWW: a collection of interconnected *documents*, linked by hyperlinks and URLs.
  - Accessible via the Internet
  - Enables computer users to locate and view multimedia-based documents
- The WWW is accessible via the Internet, like many other Internet services including e-mail, file sharing

Brief Internet History

- Late 1950s - The USSR launched the Sputnik, a series of unmanned satellites
  - Spurred the US to create the Advanced Research Projects Agency (ARPA) in February 1958 to regain a technological lead.
- 1969 - the US Department of Defense commissioned **ARPANET** for research into networking.
  - The first node was at UCLA, closely followed by nodes at Stanford Research Institute, the University of California at Santa Barbara, and the University of Utah.
- 1973 - ARPANET linked 40 machines and had international connections to England and Norway.
- January 1, 1983 - The first TCP/IP wide area network was operational when the US National Science Foundation (NSF) constructed a university network backbone that would later become the NSFNet.
  - This date is held by some to be technically that of the birth of the Internet.
  - By 1990 there were over 300,000 host computers.

… Brief Internet History

- 1995 -NSFNET was "defunded" and restrictions were lifted on commercial use, setting the stage for exponential growth in Internet usage.
  - NSFNET funding was redistributed to regional networks to help purchase Internet connectivity from the now numerous, commercial network service providers.
- 1995 - 1997 the number of sites increased by over 6 million per year to nearly 20 million host sites.
- As of January 22, 2007, 1.0935 billion people use the Internet according to Internet World Stats.
  - http://www.internetworldstats.com/stats.htm

Common Uses of the Internet

- The WWW
  - Provides instant access to a vast and diverse amount of online information using search engines
- Remote Access
  - Allows computer users to connect to other computers and information stores easily, wherever they may be across the world.
  - Provides new opportunities for working from home, collaboration and information sharing in many industries.
  - An employee on a business trip can open a remote desktop session into his normal office PC using a secure Virtual Private Network (VPN) connection via the Internet.
- Collaboration
  - The low-cost of collaborative software (.eg., email, calendaring, text chat, wiki) makes collaboration easier
  - Internet 'chat' systems allow colleagues to stay in touch in a very convenient way
- File Sharing
  - Using e-mail, a Web server, FTP server, etc
- Streaming Media
  - Many existing radio and television broadcasters provide Internet 'feeds' of their live audio and video streams (for example, the BBC).
- Voice Telephony (VoIP)
  - Provides cheap (sometimes free) Internet-based telephone calls, especially over long distances and especially for those with always-on ADSL or DSL Internet connections.

Brief History of the WWW

- 1989 - First proposal for the WWW was made
  - By Tim Berners-Lee, a scientist at CERN (European centre for High Energy Physics – Geneva)
- 1990 – first browser/editor program
- 1991 – an early WWW system released to the high energy physics community via the CERN program library
  - Including a browser, Web server and a library
  - First Web server in the US came online in December 1991 at Stanford Linear Accelerator Center (SLAC) in California
- 1993 – First version of the Mosaic browser (running on X Window System environment) released
  - At the National Center for Supercomputing Applications (NCSA), university of Illinois
  - Late 1993: over 500 known Web servers, WWW accounted for 1% of Internet traffic (the rest was remote access, e-mail and file transfer)
- 1994 – "Year of the Web"
  - World's First International WWW conference at CERN in May (about 400 users and developers in attendance)
  - Web stories got into the media
  - Second conference in USA with 1300 people in attendance (organized by CERN in October)
  - End of 1994: 10,000 Web servers (2,000 of which were commercial), 10 million users

… Brief History of the WWW

- Jan 1995 – the International WWW Consortium (W3C) was founded
  - Develops interoperable technologies (specifications, guidelines, software, and tools) to lead the Web to its full potential.
- Other developments
  - 1995 – JAVA source code was released
  - 1998 – Google was founded

Statistics

- Number of Hosts advertised in the DNS (source: http://www.isc.org/)
  - Jan 2007: 433,193,199
  - Jul 2007: 489,774,269
- In the September 2007 survey we received responses from 135,166,473 sites (source: http://news.netcraft.com/)

Language

- Top Ten Languages Used in the Web (source: http://www.internetworldstats.com)

Top Ten Languages Used in the Web
(Number of Internet Users by Language)

TOP TEN LANGUAGES IN THE INTERNET	% of all Internet Users	Internet Users by Language	Internet Penetration by Language	Language Growth in Internet (2000 - 2007)	2007 Estimated World Population for the Language
English	31.2 %	365,893,996	17.9 %	157.7 %	2,042,963,129
Chinese	15.7 %	184,001,513	13.6 %	469.6 %	1,351,737,925
Spanish	8.7 %	101,539,204	22.9 %	311.4 %	442,525,601
Japanese	7.4 %	86,300,000	67.1 %	83.3 %	128,646,345
French	5.0 %	59,207,849	15.3 %	385.4 %	387,820,873
German	5.0 %	58,981,592	61.1 %	112.9 %	96,488,326
Portuguese	4.0 %	47,326,760	20.2 %	524.7 %	234,099,347
Korean	2.9 %	34,120,000	45.6 %	79.2 %	74,811,368
Italian	2.7 %	31,481,928	52.9 %	138.5 %	59,546,696
Arabic	2.5 %	28,782,300	8.5 %	940.5 %	340,548,157
TOP TEN LANGUAGES	**85.0 %**	997,635,142	19.3 %	203.7 %	5,159,187,766
Rest of World Languages	**15.0 %**	175,474,783	12.4 %	440.3 %	1,415,478,651
WORLD TOTAL	**100.0 %**	1,173,109,925	17.8 %	225.0 %	6,574,666,417

Hypertext

- HyperText Markup Language (HTML) is the predominant markup language for the creation of web pages
- User interface paradigm
- Documents with hyperlinks
- Typed links
  - a link to another document or part of a document that includes information about the *character* of the link.
  - Useful, if only were they used
- Transclusion
  - Client-side vs. server-side includes
- Hypermedia – hypertext + multimedia
- Usually client/server architecture

Browser Implementations

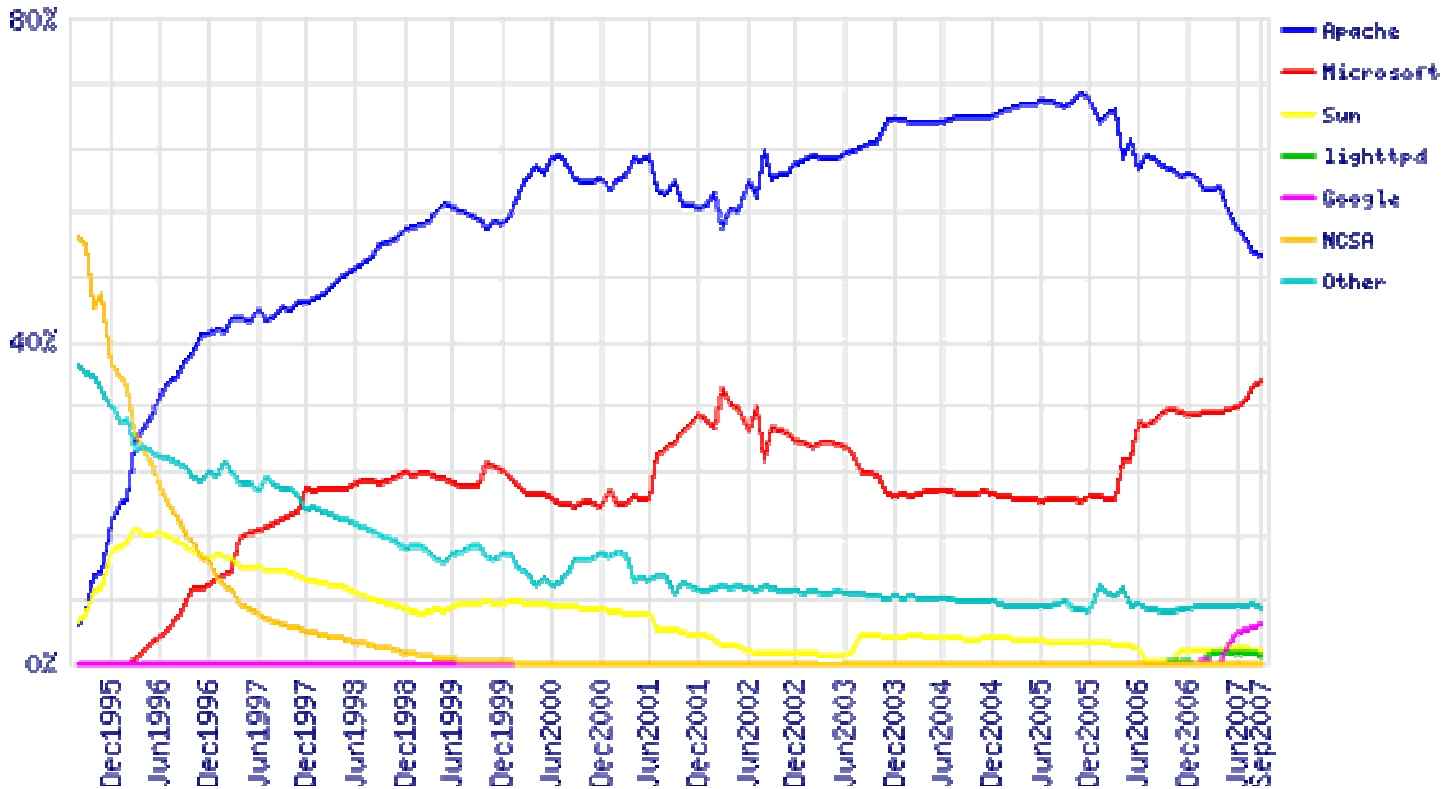
- Mosaic (1993)
- Netscape Navigator (1994)
- MS Internet Explorer (1995)
- Mozilla (1998, 2002)
- Lynx, Opera, Amaya, Safari, Konqueror…
- Firefox (2004)
- Browser wars:
  - the competition for dominance in the web browser marketplace.
    - For details see, http://en.wikipedia.org/wiki/Browser_wars

Server Implementations

- NCSA HTTPd
  - Development was suspended in 1998
- Apache = “a patchy” server
- MS Internet Information Server
- Many others in various packages
  - Many more embedded
- Server wars
  - Is Microsoft winning the Web server war?
    - see http://www.raiden.net/?cat=2&aid=287

Servers Market Share

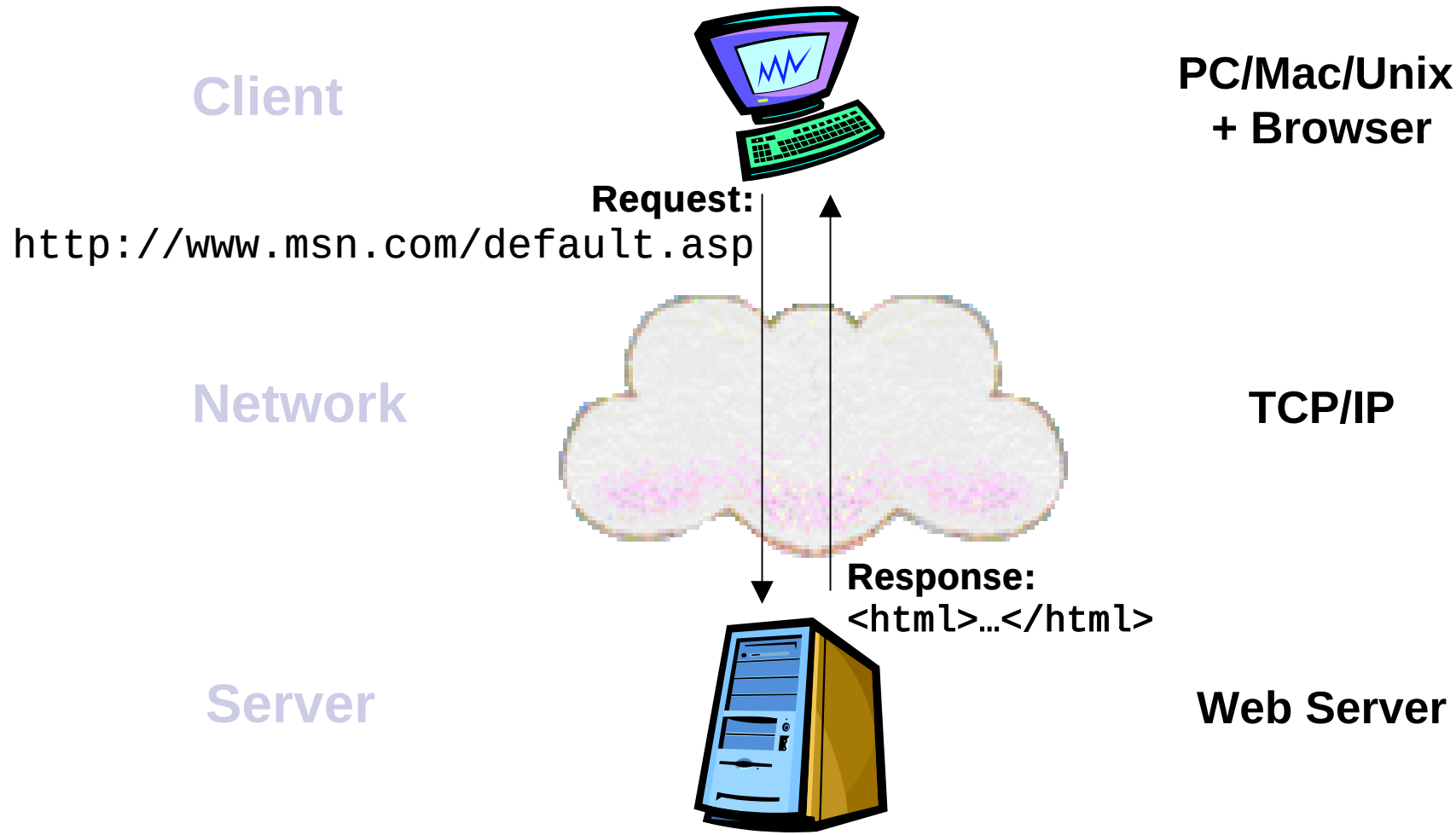
- Market Share for Top Servers Across All Domains August 1995 - September 2007 (source: http://news.netcraft.com/)



Invisible Web

- A.k.a. deep web
- Pages generated from databases
  - Parameters from forms
  - May not even have a direct URI
- Old example: product catalogues
- Increasingly becoming visible through “internal ad links” and indexes
- Estimated 500x bigger than visible web
- New example: Google search results

WWW Architecture



WWW Architecture

- Client/Server, Request/Response architecture
  - You request a Web page
    - e.g. `http://www.msn.com/default.asp`
    - HTTP request
  - The Web server responds with data in the form of a Web page
    - HTTP response
    - Web page is expressed as HTML
  - Pages are identified as a Uniform Resource Locator (URL)
    - Protocol: `http`
    - Web server: `www.msn.com`
    - Web page: `default.asp`
    - Can also provide parameters: `?name=Leon`

Proxy Servers & Firewalls

- Proxy Server
  - A server that sits between a client (running a browser) and the Internet
  - Improves performance by caching commonly used Web pages
  - Can filter requests to prevent users from accessing certain Web sites
- Firewall
  - A server that sits between a network and the Internet to prevent unauthorized access to the network from the Internet

Standardization

- Interoperability between implementations
  - How was the web before standards?
  - No more “best viewed with …”
- Stability of specification
- Industry has to agree on a standard
- Usually a compromise solution
- Best started with one candidate

Standardization Bodies

- Organizations with support from industry
- Standards not necessarily free
- ISO
- ANSI, Ecma Int'l
- Domain-specific bodies
  - Internet Engineering Task Force (IETF)
  - World Wide Web Consortium (W3C)
  - OASIS, WS-I for Web Services

Web Standards

- Governing body for Internet since 1992
  - http://www.isoc.org
- Internet Engineering Task Force (IETF)
  - http://www.ietf.org/
  - Founded 1986
  - A large open international community of network designers, operators, vendors, and researchers concerned with the evolution of the Internet architecture and the smooth operation of the Internet
  - It is open to any interested individual
- World Wide Web Consortium (W3C)
  - http://www.w3.org
  - Founded 1994 by Tim Berners-Lee
  - an open forum of companies and organizations with the mission to lead the Web to its full potential
  - W3C has around 450 Member organizations from all over the world
  - Publishes technical reports and recommendations
  - The rule-making body of the Web is the W3C
  - W3C puts together specifications for Web standards
  - The most essential Web standards are HTML, CSS and XML

Standardization of HTML

- IETF RFC 1866 – 2.0 – Nov 1995
- Then W3C Recommendations
- Also ISO/IEC standard
- 3.0 a failure
- 3.2 included changes from browsers
- 4.01 latest, made into XHTML 1.0
  - Going from SGML to XML
- XHTML 1.1 modularized

Web Browsers

- Client-side applications
- Request HTML from Web server and render it
- Popular browsers:
  - Internet Explorer
  - Netscape
  - Opera
  - others
- Also known as User Agents

Web Browsers: Compatibility Problems

- There are literally hundreds of web browsers in use around the world.
  - All of them implement the W3C document standards a little differently.
- The most commonly used browsers are Internet Explorer, Netscape Navigator, Firefox and Opera.
  - Each implements HTML, JavaScript and Cascading Style Sheets (CSS) a little differently.
  - Differences range from cosmetic to those that make Web pages look totally different
  - Each browser is free to implement "enhancements" to the W3C standard version of each of these formats.
- There are typically different flavors of the same browser type which may not be compatible
- The underlying operating systems also create difference in how the computer displays graphical elements and text differently.
- HTML editors are, on their part, notorious for creating non-compliant and garbage code.
- A cross-browser compatible Web page will look more or less the same in all of the existing Web browsers

Designing for Cross-Browser Compatibility

- Obviously, 100% compatibility with all potential browsers is impossible.
- Write clean code that conforms to the W3C standards to get consistent results across all browser platforms.
- Write your code by hand, e.g., using notepad
  - If you must use a HTML editor, the best choice for compatibility is Dreamweaver and worst is FrontPage.
- Use code cleaners and validators (freely available)
  - http://tidy.sourceforge.net/
  - http://validator.w3.org

1.2 HTTP and HTTPS

- Introduction
- Request/Response types
- MIME Data Formats
- HTTPS

Hypertext Transport Protocol (HTTP)

- HTTP is an application protocol based on client-server architecture, designed for delivering hypermedia information on the web.

- The design goals of HTTP are:
  - light protocol: not consuming too much resources
  - fast protocol: need to retrieve many widely distributed documents as fast as possible

- HTTP evolved into multiple, mostly backward-compatible protocols versions
  - HTTP 1.0: simple
  - HTTP 1.1: more complex

- For full details about the HTTP protocols, refer to the links: http://www.ietf.org/rfc/rfc1945.txt (1.0) and http://www.ietf.org/rfc/rfc2616.txt (1.1).

...HTTP

- HTTP is a stateless protocol
  - Doest not retain information about users between requests
  - Each HTTP request is independent of previous and subsequent requests
  - State information can be maintained using cookies
- *Persistent* (or *keep-alive*) connections
  - Connections that allow more than one request/response per TCP/IP connection
  - Only work well when not using proxy servers
  - Used for efficiency purpose

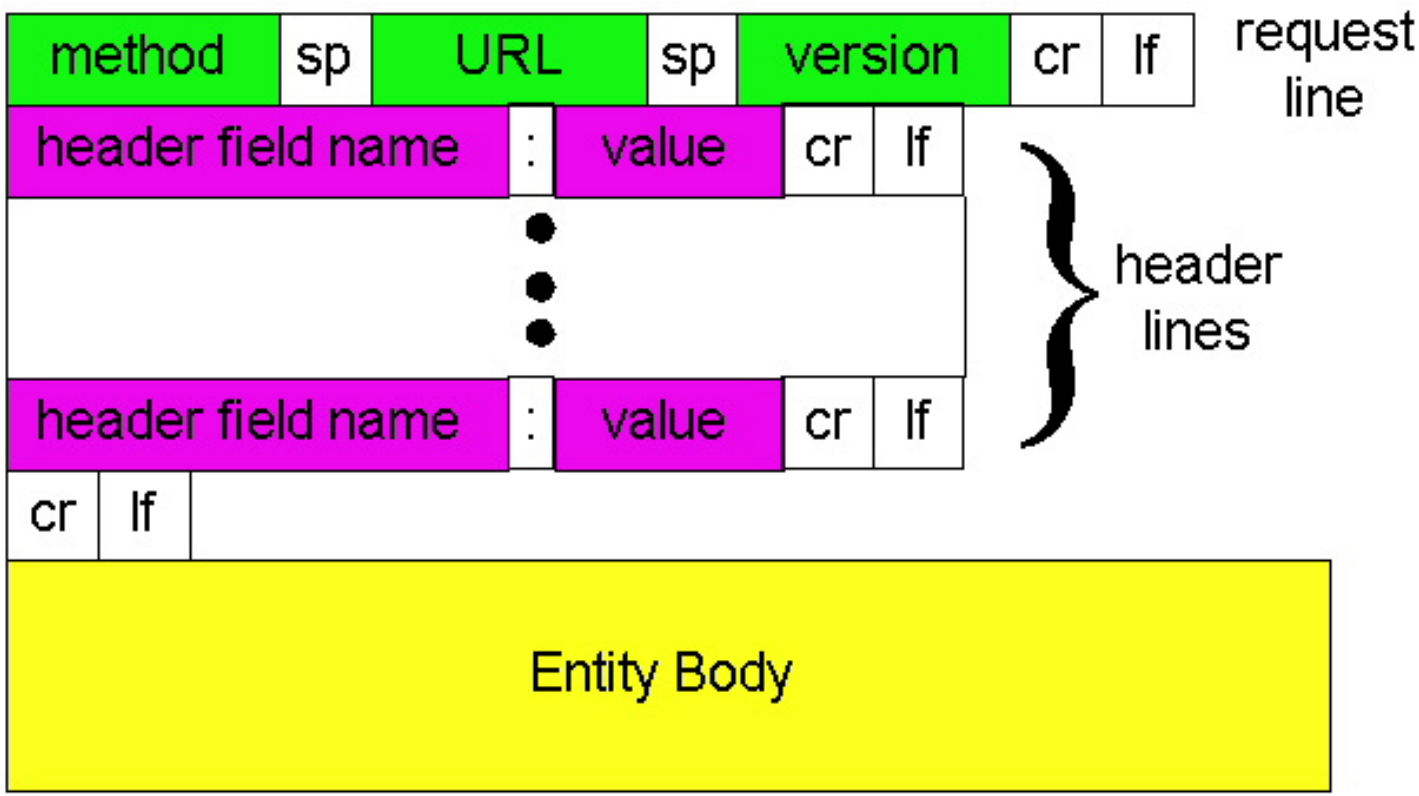
…HTTP

- HTTP 1.0
  - Persistent connections not allowed by default; allowed only when explicitly negotiated
- HTTP 1.1:
  - Persistent connections allowed by default
  - Works well with proxies
- A client tells in the beginning of a request the HTTP version it uses, and the server uses the same or earlier version in the response

HTTP Message Format

- The format of request and response messages are similar
- Both consist of
  - An initial line (different in both)
  - Zero or more header lines
  - A blank line
  - An optional message body
- The request/response line and each header line must end with CRLF (“\r\n”). The Request body is sent in binary format.

HTTP Request Format



HTTP Request

- The request line has three parts, terminated by CRLF:

  Method Request-URI HTTP-Version CRLF

- Methods (case sensitive) include:
  - **GET**: request document named by request-URI. Any parameters for the request are appended to the request-URI
  - **HEAD**: request only header information about request-URI
  - **POST**: similar to GET but request parameters are provided through the Message Body.

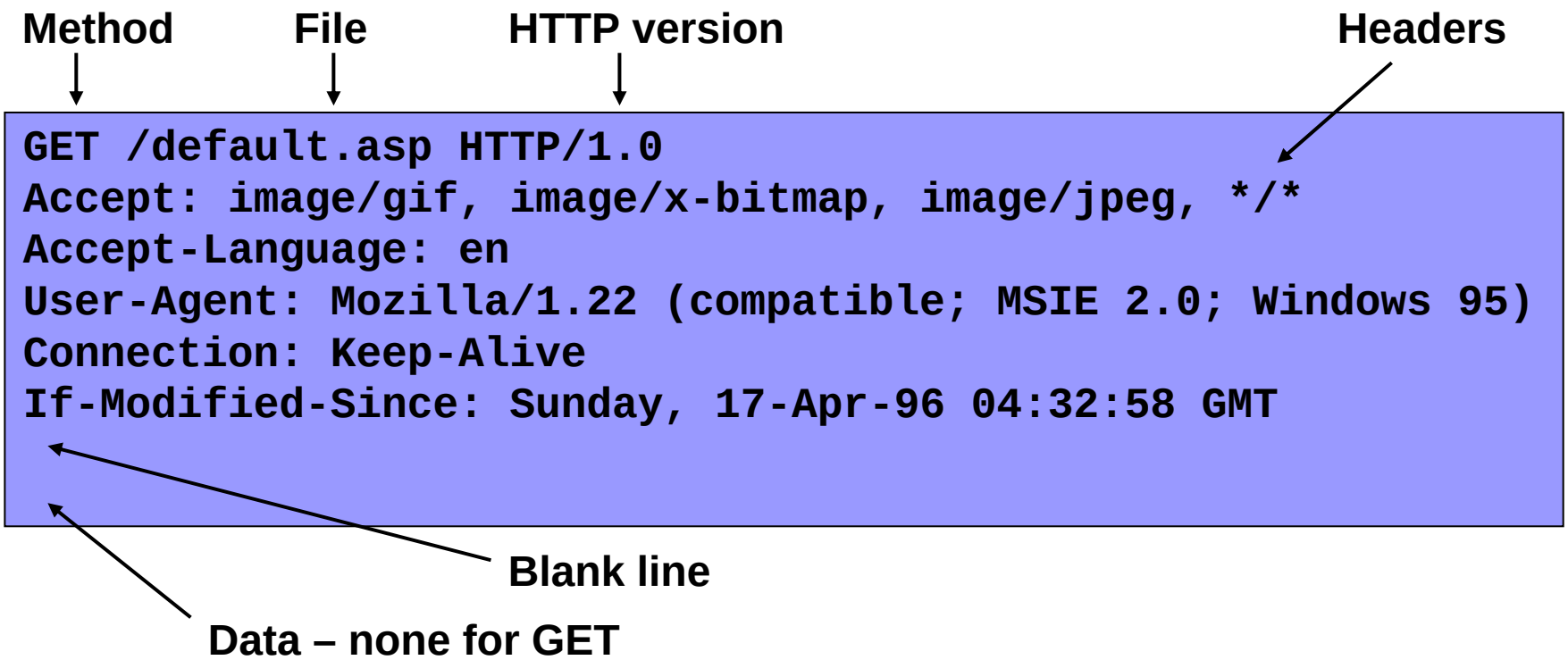
- Request-URI specifies the full path of the resource being requested and must begin with "/":
  - e.g: /swe344/lectures/lecture1.html
  - **Note**: URI stands for Universal Resource Identifier – superset of URL and URN

- HTTP-version specifies the version of HTTP of the client making the request. Values are: HTTP/1.0 or HTTP/1.1

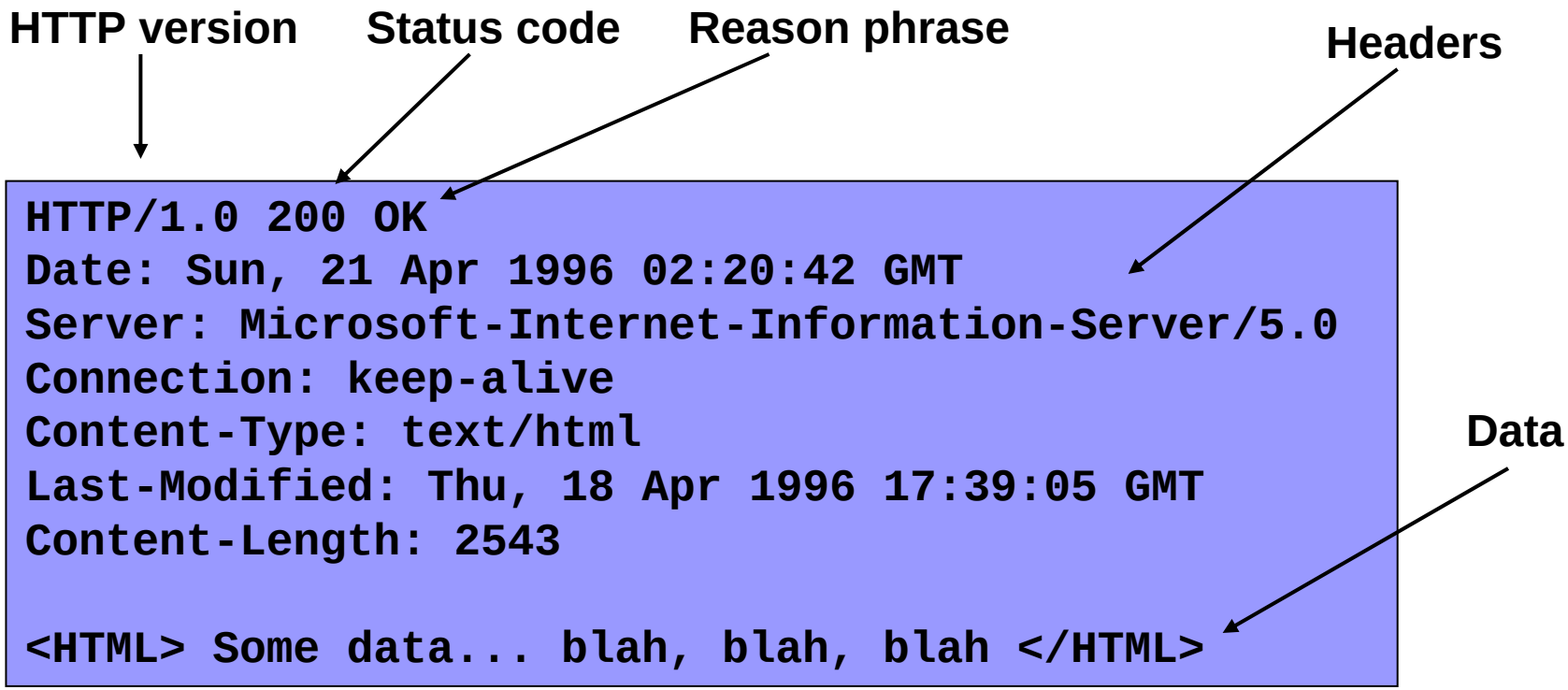
HTTP Headers

- Headers are used by both the Client making a request or by the Server responding to the request.
  - Headers provide information about the request, response, object being requested/sent, server or client.
- The headers have the form "Header-Name: value", ending with CRLF.
  - The header name is not case-sensitive (but the value may be).
  - Any number of spaces or tabs may be between the ":" and the value.
- Some common header names:
  - Accept: what format is acceptable (client)
  - Content-Length: length of the message (client/server)
  - Content-Type: type of the content (server)
  - Date: date sent (sever)
  - Expires: expiry date of the content (server)
  - Last-Modified: Last modification date (server)

HTTP Request



HTTP Response



- A response also consists of four parts but the first line is called the status line and has three parts:

HTTP-Version **Status-Code** **Status-Text** CRLF

HTTP Response

- HTTP-Version specifies the HTTP of the server responding to the request: HTTP/1.0 or HTTP/1.1
- Status-code is a three-digit integer indicating the status of the request.
- Status-Text explains the status.
- First digit identifies the category of the response
  - **1xx** indicates an informational message only
  - **2xx** indicates success of some kind
  - **3xx** redirects the client to another URL
  - **4xx** indicates an error on the client's part
  - **5xx** indicates an error on the server's part
- The most common status codes are:
  - **200 OK** : The request is successful
  - **404 Not Found** : The requested resource doesn't exist.
  - **500 Server Error** : An unexpected server error.

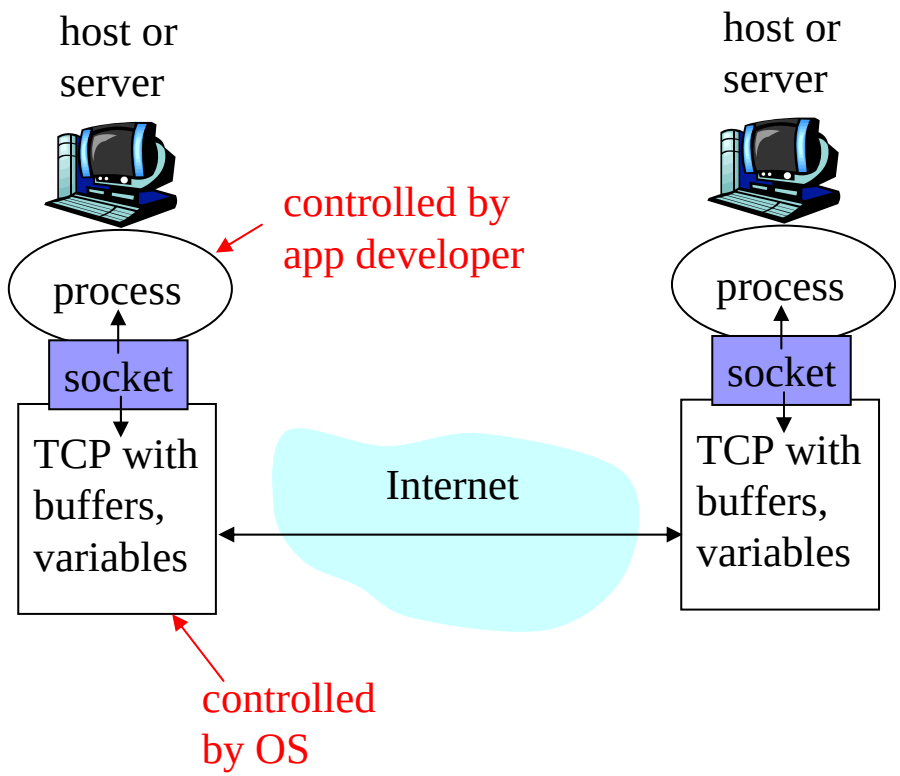
HTTP Server Status Codes

Code	Description
200	OK
201	Created
301	Moved Permanently
302	Moved Temporarily
400	Bad Request – not understood
401	Unauthorized
403	Forbidden – not authorized
404	Not Found
500	Internal Server Error

- 401: Header specifies the authorization scheme needed. So, request must be made with authorization.
- 403: Authorization will not help as the page is forbidden.

Communicating Across Network

- Process sends/receives messages to/from its socket
- Socket analogous to door
  - sending process shoves message out door
  - sending process assumes transport infrastructure on other side of door which brings message to socket at receiving process



HTTP Transaction Example

1. Telnet to your favorite Web server:

```
telnet www.ccse.kfupm.edu.sa 80
```

Opens TCP connection to port 80 (default HTTP server port) Anything typed in sent to port 80 at **www.ccse.kfupm.edu.sa**

2. Type in a GET HTTP request:

```
GET /~sahalu/index.html HTTP/1.0
```

By typing this in (hit carriage return twice), you send this minimal (but complete) GET request to HTTP server

3. Look at response message sent by HTTP server!

Cookies

- A mechanism to store a small amount of information (up to 4KB) on the client
- A cookie is associated with a specific web site
  - Enables a Web server distinguish between clients
  - Used to customize pages
- Cookie is sent in HTTP header
- Cookie is sent with each HTTP request
- Can last for only one session (until browser is closed) or can persist across sessions
- Can expire some time in the future

Multipurpose Internet Mail Extensions (MIME) Types

- HTTP requires that data be transmitted in the context of e-mail-like messages, even though the data may not actually be e-mail.
- An Internet Standard that extends the format of e-mail to support text in character sets other than US-ASCII, non-text attachments, multi-part message bodies, and header information in non-ASCII character sets.
- A standard for specifying the format of content
- Helps browsers determine how to display the data
- application/*
- audio/*
- image/*
  - image/jpeg
  - image/tiff
- video/*
  - video/quicktime
  - video/mpeg
  - video/x-msvideo
- text/*
  - text/xml
  - text/rtf
  - text/html
  - text/plain

Pages with Multiple Types

- Each entity (ex. image) is standalone HTTP request
  - Page with many pictures creates many connections
- Each response therefore has appropriate MIME settings

HTTPS

- A secure version of HTTP with a different default port (443) and an additional encryption/authentication layer between HTTP and TCP.
  - Syntax: https://
- Invented by Netscape Communications Corporation to provide authentication and encrypted communication
  - Widely used on the Web for security-sensitive communication such as payment transactions and corporate logons.
- Strictly speaking, https is not a separate protocol
  - Refers to the combination of a normal HTTP interaction over an encrypted Secure Sockets Layer (SSL) or Transport Layer Security (TLS) transport mechanism.
  - Ensures reasonable protection from eavesdroppers and man-in-the-middle attacks.
- The level of protection depends on
  - the correctness of the implementation by the web browser and the server software and
  - the actual cryptographic algorithms supported.

1.3 Internet Search

- The following three methods of reaching information on the Web can be identified:
  - Search Engine
  - Directories
  - Portal

Introduction to Internet Search

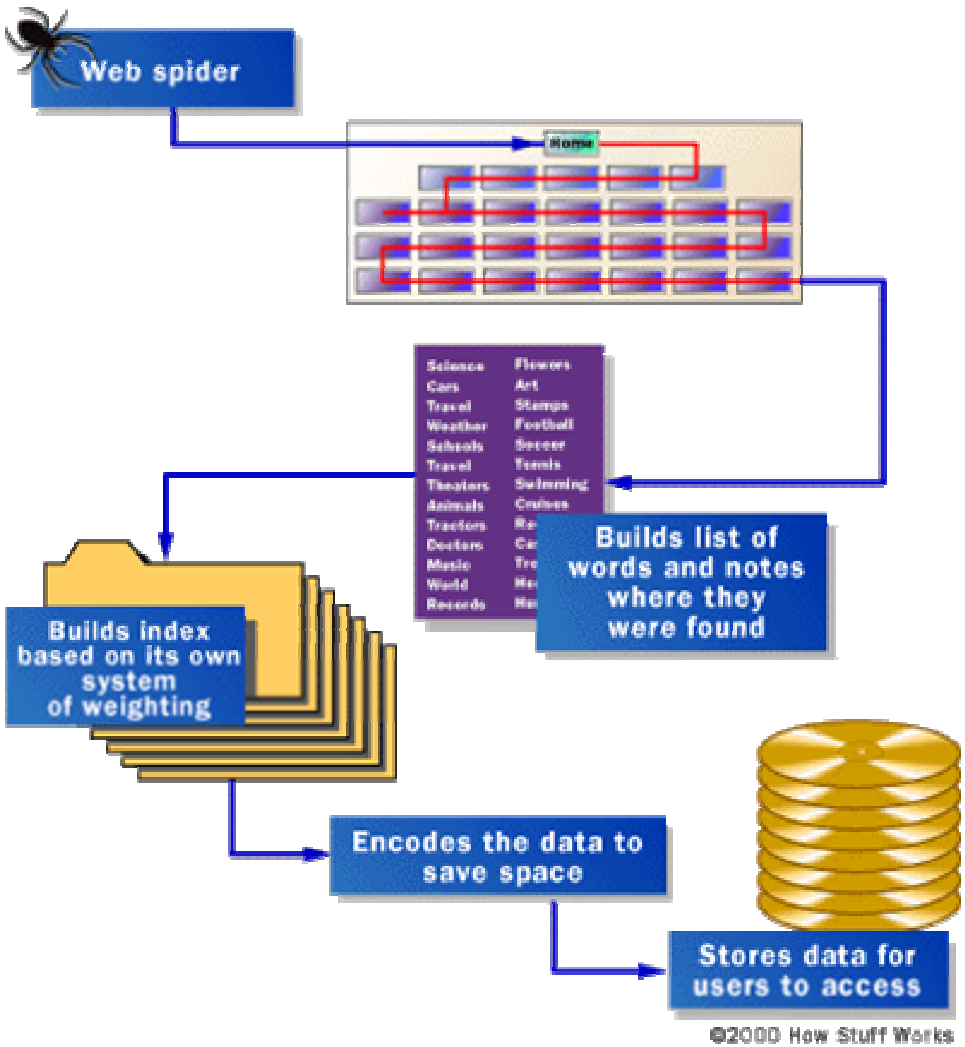
- A **search engine** is an information retrieval system designed to help find information stored on a computer system, such as on the Web

- The very first tool used for searching on the Internet was Archie
  - Created in 1990 by Alan Emtage, a student at McGill University in Montreal.
  - Archie downloaded the directory listings of all the files located on public anonymous FTP sites, creating a searchable database of filenames
  - However, Archie could not search by file contents.

How Search Engines Work

- A **search engine** works in the following sequence:
  1. Web crawling
  2. Indexing
  3. Searching
- Web search engines work by storing information about a large number of web pages, which they retrieve from the WWW itself.
- These pages are retrieved by a Web crawler (sometimes also known as a spider)
  - Crawling starts with a popular Web site containing lots of links, such as Yahoo
  - Crawling continues until it finds a logical stop, such as a dead end with no external links or reaching the set number of levels inside the Web site's structure
- The contents of each page are then analyzed to determine how it should be indexed
  - For example, words are extracted from the titles, headings, or special fields called meta tags.
    - See a sample analysis: http://en.wikipedia.org/wiki/Inverted_index
- Data about web pages are stored in an index database for use in later queries.
  - Early engines held an index of a few hundred thousand pages and documents, and received maybe one or two thousand queries a day
  - Today, a top search engines will index hundreds of millions of pages, and respond to tens of millions of queries a day

Standard Web Search Engine Architecture



… How Search Engines Work

- When a user makes a query, typically by giving key words, the engine looks up the index and provides a listing of best-matching Web pages according to its criteria
  - usually with a short summary containing the document's title and sometimes parts of the text
- The usefulness of a search engine depends on the relevance of the **result set** it gives back
- Most search engines employ methods to rank the results to provide the "best" results first.
- Most Web search engines are commercial ventures supported by advertising revenue
  - Some employ the controversial practice of allowing advertisers to pay money to have their listings ranked higher in search results
- Those who don't accept money for their search engine results make money
  - by running search related ads alongside the regular search engine results.
  - everytime someone clicks on one of these ads.

Challenges Faced by Search Engines

- Size of the Web
  - Contains more than 3 billion documents, growing very fast and not indexed in any standard vocabulary
- Currency
  - Many Web pages are updated frequently, which forces the search engine to revisit them periodically.
- Relevancy
  - Because the queries one can make are currently limited to searching for key words, may result in many false positives
  - Better results might be achieved by using a proximity-search option or using organic search engines.

… Challenges Faced by Search Engines

- Problem with dynamically-generated Web sites
  - Because these sites may be slow or difficult to index, or may result in excessive results, perhaps generating 500 times more Web pages than average.
- Search engines can be tricked
  - To return pages, in favor of the trick makers, which contain little or no information about the matching phrases.
  - Making the more relevant Web pages pushed further down in the results list
- Indexing secured pages
  - Content hosted on HTTPS URLs pose a challenge for crawlers which either can't browse the content for technical reasons or won't index it for privacy reasons.

The Invisible Web – 4 Types

1. Opaque: search engines (intentionally) choose not to index
   - Depth of crawl is limited – sometimes for cost reasons
2. The Private Web: password protected
   - robots files disallows spiders access, “noindex” meta tag prevents access
3. The Proprietary Web: registration required (either fee or free)
   - Examples: The New York Times, The Well, The Wall Street Journal Interactive Edition.
4. The Truly Invisible Web: can’t search certain file formats and databases
   - *Non-HTML/text content* - textual content encoded in multimedia (image or video) files or specific file formats not handled by search engines.
     - file formats like PDF, Flash, Shockwave, etc
   - Recently some of the commercial search engines have added image and PDF files to their indexes.

How Do I Use The Invisible Web?

- Why search the invisible Web?
  - The materials found on the Invisible Web are often more focused, current, and professionally relevant than what you can find on the public web using search engines.
- One way to explore the deep web is by using human crawlers instead of algorithmic crawlers:
- Through the Direct Search site
  - put together by Gary Price, a librarian and information research consultant.
  - nicely organized into searchable categories and is updated frequently.
- Through the Invisible Web Directory,
  - put together by the aforementioned Gary Price and search guru Chris Sherman.
  - a directory of searchable databases, organized by subject.
- The Virtual Library is simple and easy to use, with annotated subject links.

Examples of Invisible Web Sites

- Dictionaries http://www.m-w.com
- Telephone Numbers http://www.infospace.com
- Clinical Trials http://www.clinicaltrials.gov
- Library Catalogs http://www.libdex.com/webcats
- Philanthropy and Grant Information http://lnp.fdncenter.org/finder
- Translation Tools http://world.altavista.com

Major Search Engines

- Google (http://www.google.com/)
  - Try the Googlewhacking game: type two words for Google search in the hopes of receiving *exactly* one result!
- AltaVista (http://www.altavista.com/)
- Alltheweb (http://www.alltheweb.com/)
- Kartoo (http://www.kartoo.com)
- Teoma (http://www.teoma.com)
- Vivisimo (http://www.vivisimo.com)
- Why does the same search on different search engines produce different results?

Directories

- A **web directory** is a repository or database of information on the Web
  - As opposed to a conventional database, a directory is heavily optimized for reading, with the assumption that data updates are very rare compared to data reads.
  - Commonly, a directory supports search and browsing in addition to simple lookups.

- A web directory is not a search engine, and does not display lists of web pages based on keywords
  - A directory lists web sites by category and subcategory.
  - A whole web site, rather than one page or a set of keywords, often limited to inclusion in only one or two categories.

- Directories have various types of listings, often dependant upon the price paid for inclusion:
  - Free Submission - there is no charge for review of the site
  - Reciprocal Link - the site submitted must link back to the directory in order to be listed
  - Paid Submissions - a fee is charged for reviewing the submitted link
  - No Follow - there is a rel="nofollow" attribute associated with the link, meaning search engines will not follow the link.
  - Featured Link - the link is given a premium position in the category where it is submitted
  - Featured Homepage Link - the link may be listed on the homepage of the directory.

Who Creates Directories?

- Libraries
- Nonprofit organizations
- Universities
- Dot-Com businesses
  - but they are probably portals too
- Many directories, including the Open Directory Project and the World Wide Web Virtual Library, are edited by volunteers, who are often experts in particular categories.

A Sampling of Directories

- Ansearch
  - Web search and Directories focusing on the US, UK, Australia and New Zealand.
- Best of the Web Directory
  - Lists content rich, well designed websites categorized both by topic and by region.
- Open Directory Project (aka DMoz or ODP)
  - The largest directory of the web. Its open content is mirrored at many sites, including the Google Directory.
- World Wide Web Virtual Library (VLIB)
  - The oldest directory of the Web.
- Yahoo! Directory
  - The first service Yahoo! offered.

Portals

- A **Web portal** is a site on the Web that typically provides personalized capabilities to its visitors, providing a pathway to other content
- Many of the portals started initially as either web directories (notably Yahoo!) and/or search engines (Excite, Lycos, AltaVista, infoseek, and Hotbot among the old ones).
- Portals offer a one-stop shopping look
- Portals include e-mail, chat, auctions, news, weather, horoscopes, stock info, and more.
- Portals want to be YOUR starting point

A Sampling of Popular Portals

- Yahoo! : www.yahoo.com
- Portals to the World from the Library of Congress: www.loc.gov/rr/international/portals.html
- AltaVista: www.altavista.com

Directories Vs Search Engines

- When should you use a directory?
  - When you have a broad topic
  - When you want experts to recommend sites
  - When you want to avoid irrelevant sites
  - Examples topics:
    - Disabilities
    - Civil War
    - Welfare

Directories Vs Search Engines

- When should you use a search engine?
  - When you have a narrow topic
  - When you are looking for a specific website
  - When you want to search for a file type or language
  - Examples:
    - Americans with Disabilities Act
    - Battle of Gettsyburg
    - Welfare to Work
  - Good For: Precision searches, using named people or organisations, searching quickly and widely, topics which are hard to classify
  - Not Good For: Browsing through a subject area