Studies (How Stuff Works) Acronym Finder Alex Catalogu Bartleby.com: Strunk's Dictionary.com Texts Element of Style (1918). The Library of Congress
The Weather Channel 7 (The Online The Online Books Page http://www.thesaurus.com/ Internet Public Library: Books Merriam-Webster Online Encyclopedia.com rld Factory Biography.com MSNBC **Washington Post** The Old Farmer's Almanac USA Today Google Image Search Yahoo NewsNew Los Angeles Times Web Search Home Page (Texi2html's Homepage) MetaCrawler Welcome to Base21 Violations of the GPL, LG For datio (III)

and GFDL - GNU Project Free miniB

MetaPad

MetaPad

Violations of the GPL, LG For datio (III)

Lew School Reverse School Reverse School Reverse School Reverse Rev Blogarama Why There Are Not GI on GNU Pages Aph An introduction to practical computing of the Dougal's Slightly ns DeedDougal's Slightly Less Funky Aboutmedi Homepage WebMask help - Contents (Italy Paper Money INDEX) http://ca.wikipedia.org/ Linux Today - Linux News On bletype.org : Get Movable) Montana Local i M clana Internet Time. Personal Singapore Travel and Hotel Hotels, Montana Late **Boing Boing** Alex Kinge Little's Guide Infoportal Infoportal ina prima - Wikipedia - Wikipedia Wikimedia Foun free encyclopedia Minnesota Encyclopedia : Maps Iberghi Italia Movable Type Journalized Weather - Travel - History er sa su Stashdot 🗐 ine Register Friendly WordPress 118#39;870 botals Bollywood and the All India Level Double-A Conformance to - Hauptseite - Wikipedia Mwaisse pä Ydje - Wikipe Mozilla Firefox Web Content Accessibility Kurobi... or carical Encyclopedia: Infoportal **Guidelines** à sà Zà sy⊓à ray⊓pedia O Artspaci > Traver & Hoter Guide > Online Infoportal Portada - Wikcionario http://validator.w3.org/check?/illa.org/ uri=referer Weather > Tom's Hardware G(http://validator.w3.org/check/ com: The Source for Level A Conformance to Web (Esileht - Vikipeedi // (Tom's (ã □ iã □ : ã □ ³ã □ □ã □ %ã □ - Wikipedia) ::: Welcome to KikuMobile.com Asinah.net Infoportal W3C CSS Validator Infoportal ļ: Main Dana - Wikipedie PAigina principal - HlavnÃ- strana - Wikipedie Content Accessibility http://jigsaw.w3.org/css-valid HTML Validation Guidelines 1.0 http://si.wikipedia wikibooks portal - wiki Main Page - Wikipedia ipedia ain Page (Glavna stran - Wiki edija) ator West Virginia Encyclopedia : Bok Viewable With Ann Browser World Wide Web Consortium Maps - Weather - Travel -History Hafan - Wicipedia **GNU General Public License** Main (Hoofdpagina - Wikipedia NL)edia (GPL) A List Apart Infoportal Unang Pahina - Wikipedia Free Software (SourceForge)
Infoportal GNU Free Documentation License Inter//septit.wikipedia.org/wik W3C HTML Home Page Infoportal Web Standards Project nt Accessibility
Guidelines 1.0 Definition World Wide Web Infoportal OpenOffice.org **Geschenk** Informationen zu Infoportal infoportal **Kontaktlinsen HTTP/1,0 200 OK Infoportal

Infoportal The K Desktop Environment Linux.org Opera Software Infoportal Welcome to SUSE LINUX

CAFFEEOD/ CLIES MACIE

Library of Congress: Country

Hypertext

Hypertext

- Text with links
- Allows free association
- More creative learning

Vannevar Bush

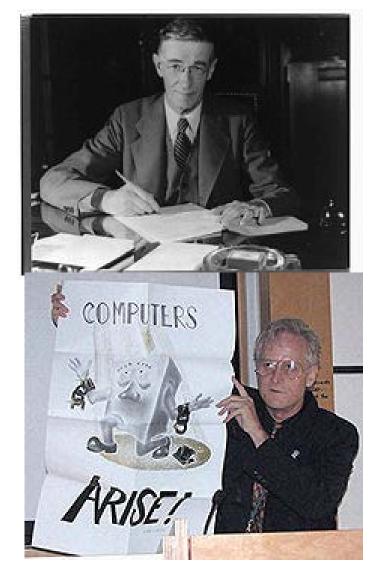
 1945 Essay describing, among others, the MEMEX system

Ted Nelson

- 1960 Xanadu project
- First computer hypertext system

• Tim Berners-Lee

1989 starts the WWW project at CERN

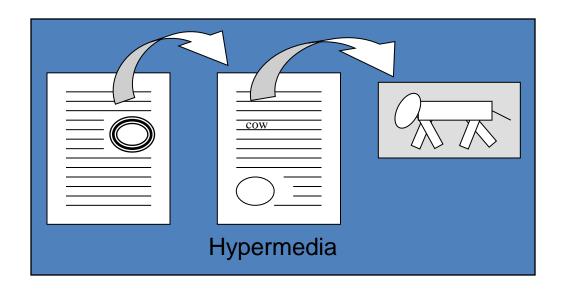


http://en.wikipedia.org/wiki/Www

Multimedia and Hypermedia

Multimedia

- The integration of many forms of media
- Text
- Images
- Sound
- Animation



Hypermedia

- The combination of Hypertext and Multimedia
- Links are made between any media (Hyperlinks)

The WWW project

Background: CERN

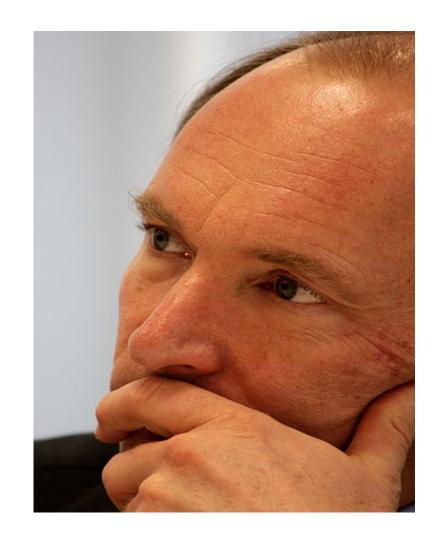
- Many networks existed
- Each network had many documents

Aims

- Access documents from any network in seamless manner
- World-Wide (distributed)
- Easy to add documents (dynamic)

Proposal

- Use Hypertext
- No intention to support hypermedia
- Research only



Evolution of the web (1)

- 1989 Tim Berners-Lee begins work on the WWW project
- 1991 WWW operational at CERN
- 1992 WWW goes public
- 1993 Mosaic created by Marc Andreessen (First GUI browser)
- 1994 US Senate allow commerce on Internet
- 1994 Netscape Communications formed, Yahoo! formed

Evolution of the web (2)

• 1995 Microsoft Internet Explorer

 1998 Netscape became open-source, developed into Mozilla Google founded

1991-2001 "Dot-com" boom and bust

2002-on The web becomes ubiquitous

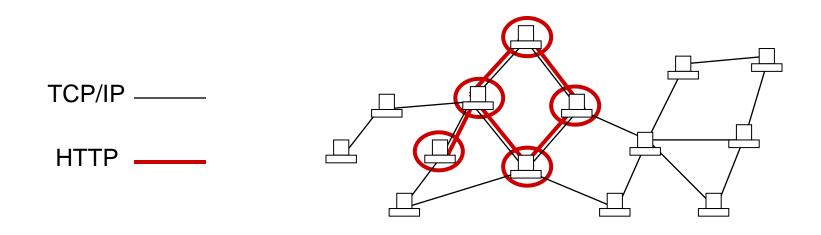
Technical Details

HTTP

- Hypertext Transfer Protocol
- Language used to transfer Hypertext documents
- Client-Server Model

WWW

Global body of information available using HTTP



Cyberspace Addresses

Uniform Resource Locators (URL)

Address used for any web resource

Protocol

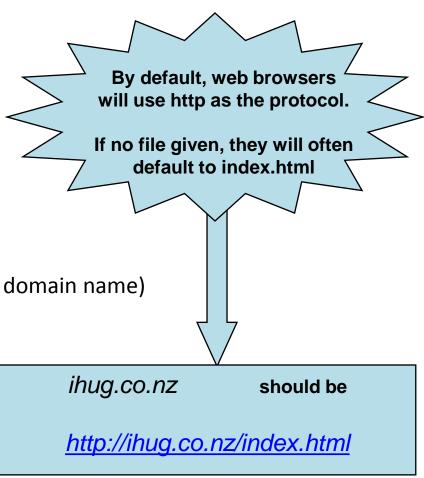
- Name of the protocol used
- ftp:// http:// https://

Domain

- Name of a host computer (IP address or domain name)
- www.cs.auckland.ac.nz

File/ Resource

- Path of the file
- /~yunsing/index.html



Terms

Web Site

 A collection of Web pages related to a single topic or theme. Normally designed and maintained by a single individual or organisation

Web Page

A hypermedia document designed for the WWW

Web Browser

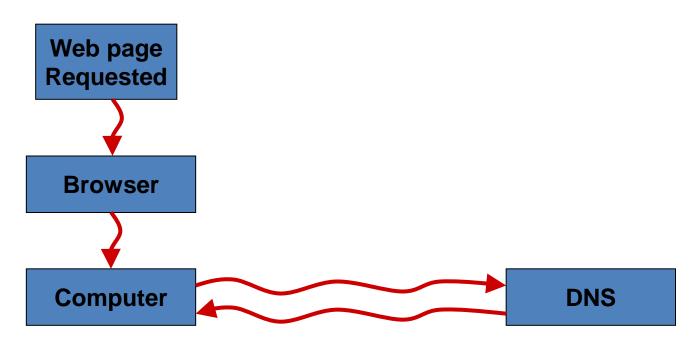
- Software used to access information on the World Wide Web
- Sends requests to a web server
- Client

Web Server

- Software that makes local files available through the web
- Fulfils requests from a web browser
- Server

Accessing a web page (1)

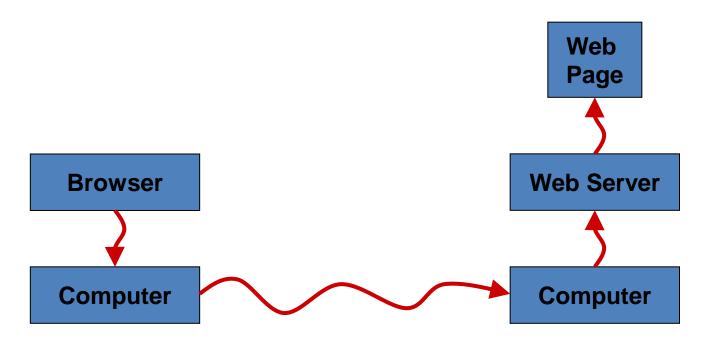
- Client (Web Browser) runs on the local machine
 - User requests a web page
 - Client contacts the DNS to find the IP address.



DNS resolves the domain name

Accessing a web page (2)

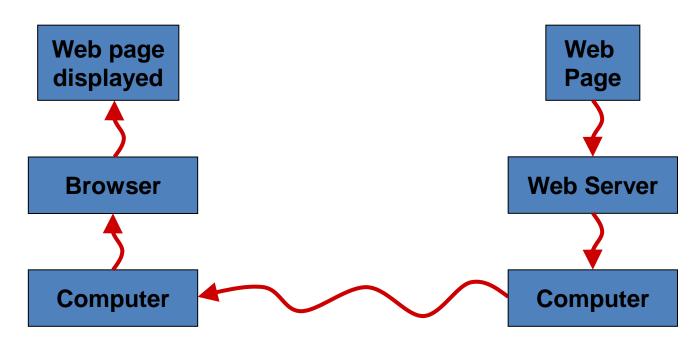
- Web server runs on the destination machine
 - Request sent to destination domain
 - Web server accepts the request and finds the web page.



Web page requested from destination domain using HTTP

Accessing a web page (3)

- Web page is sent from the server to the client
 - Client (web browser) displays the page



Web page sent from server to client using HTTP

More Terms

Proxy

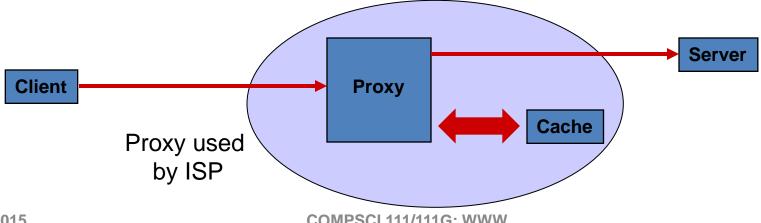
 A computer which sits between the client and server, intercepts and processes requests

Cache

- Store of information for quick access
- (e.g. caching may be used by proxy servers to speed web use)

Firewall

Prevents unauthorised access to or from a private network



COMPSCI 111/111G: WWW 3/08/2015 13

Logging web page access

Client keeps log

History in web browser

Operating System keeps log

Requests are logged by Windows on local machine

ISP keeps log

- Requests from "IP address" to "IP address" for "Page Name"
- Some ISPs may have the logs available for users to check

Web server keeps log

- Gets requests from "IP address" for "Page Name"
- Your viewing habits are being tracked!

http://en.wikipedia.org/wiki/Google_and_privacy_issues

Navigating

Finding information

- Lots of users have problem finding new information
- Lots of users have problems finding known information
- Web is very large, rapidly changing

Search Engines

- Automated
- Essential
- Our gateway to information

Problems with the structure of web pages

Broken Links

- Pages which have been moved
- Referring page is not altered

No inherent security/ tracking/ accounting system

- Difficult to have layers of security
- Forces publishers to rely on advertising revenue

No inherent information indexing

- Much of the information is not accessed by search engines
- Information created on-the-fly from databases
- Information in other formats (postscript, pdf, archived)

Search Engines

Companies (Worldwide Market Share 2014)

- Google (66.44%)
- Baidu (China) (11.15%)
- Microsoft Bing (10.29%)
- Yahoo (uses Bing since 2009) (9.31%)
- Specialised Alternatives
 - DuckDuckGo
 - Wolfram Alpha

Automatically search every web page

- Archive the contents
- Index all the words
- Try to determine the relevance of the page

http://en.wikipedia.org/wiki/Search_engines

http://www.netmarketshare.com/search-engine-market-share.aspx?qprid=4&qpcustomd=0

Can we trust the search engines?

Search Engines

- gateway to information
- pages are rated (how?)
- Since Ranking Algorithms are secret, we have to trust but can we?

Publishers/ Advertisers

- Trick the search engines (repetition of words,)
- Search engines tailor advertisements to searches
- Pay for higher rating?
- Who owns the search engines? How do they make their money?

Censorship

- 'Right to Forget' in the EU since June 2014
- But even before that many requests for deletion (DMCA, local laws etc.)

http://en.wikipedia.org/wiki/Google_censorship

Crawling the Web: Where do search engines get their information?

- A 'Web crawler' is an internet bot that systematically browses the WWW and indexes encountered websites.
- Might store encountered websites for later processing.
- Start off with list of URLs and add any links encountered on these pages to their 'To-Visit' list.

Follows a number of policies

- Selection: Only 'important' pages are indexed (2009: Large search engines index 40%-70% of indexable web, up from 16% in 1999)
- Re-visit: When should the index for what page be updated, cost vs benefit.
- Politeness: Crawlers are really good at getting lots of data quickly they have to be careful not overload a website.
- Parallelization: How do several crawlers split the task/web and recombine their results.

Searching

Searching Tips

- Learn how to use the advanced features of your search engine
- If the first page is not promising, choose different key words
- Some tips on searching with Google: http://mashable.com/2011/11/24/google-search-infographic/

Finding useful sites

- Use specialist sites for specific searches
- Build a list of useful resources
- IMDB, New York Times, Ebay, BBC
- Trademe, NZ Herald, IRD, Woolworths, GOVT

Google Top Trending 2013

NZ

- American's Cup
- Paul Walker
- Roya Baby
- Cory Monteith
- Lorde
- Wheedle
- Earthquakes Today
- iPhone 5s
- Harlem Shake
- X Factor NZ

Global

- Nelson Mandela
- Paul Walkner
- iPhone 5s
- Cory Monteith
- Harlem Shake
- Boston Marathon
- Royal Baby
- Samsung Galaxy S4
- PlayStation 4
- North Korea

Google trends: topcharts NZ

Google Trends: topcharts Global

Google Top Trending 2014

NZ

- FIFA World Cup
- Robin Williams
- Commonwealth Games
- Malaysia Airlines
- iPhone 6
- Jennifer Lawrence
- Charlotte Dawson
- Flappy Bird
- Spark
- Ebola

Global

- Robin Williams
- World Cup
- Ebola
- Malaysia Airlines
- ALS Ice Bucket Challenge
- Flappy Bird
- Conchita Wurst
- ISIS
- Frozen
- Sochi Olympics

Google trends: topcharts NZ

Google Trends: topcharts Global

Google News

- News aggregator, variation of the search engine
- Automatically searches thousands of publications and displays summaries, relevant parts. Examples:

Where in Ukraine is Viktor Yanukovych?

- Yanukovych's exact whereabouts remained unknown
- Yanukovych surfaced Saturday in the city of Kharkiv

Robots will be smarter than us all by 2029, warns Google futurologists

- computers will be able to understand our language, learn from experience
- By 2029 they will outsmart even the most intelligent humans, according to Google's director of engineering Ray Kurzweil.

Many Publishers/News Agencies unhappy

- Google reuses (snippets of) their content
- Shut down in Spain in December 2014, where new law requires payment for reuse

(Online) innovations

Voice over IP

- Google Hangouts, Skype, ISPs
- Cheap/free voice communication

Peer to Peer networks

- BitTorrent
- Swarming downloads

Wolfram

- WolframAlpha: searching = computing
- Wolfram language: knowledge-based programming

Free Books

- http://digital.library.upenn.edu/books/
- http://books.google.com

Internet for everybody anywhere

- Google Project Loon (http://www.google.com/loon/): using high-altitude balloons to create a wireless network that provides internet in rural and remote areas.
- Outernet (https://www.outernet.is/en/): free internet anywhere in the world through satellites

Web-agents and other future directions

Computer programs that operate on your behalf

- Tracks all your browsing habits
- Makes suggestions based on what you have read
- Recommender Systems: Big, active research area, permeates many areas (shopping, video streaming, search)

• TiVo Suggestions, Netflix recommendations

- Similar principle with television viewing
- Netflix Prize (2006-2009), \$1 Million: Improve Netflix' own algorithm for predicting user ratings for movies based on previous ratings by 10%.

The Internet is changing extremely rapidly

- Too fast for legislation to keep up
- Too fast to predict the future

Some things coming soon

- Wearable PC's
- Integrated Media (Interactive T.V.)
- Household Appliance connections
- And of course: new approaches to internet-related crime.