Searching the WWW

Locating the right information on the WWW requires effort.

Looking In the Right Place

Google is not the first place to look.

- Go directly to a site -- www.irs.gov
- Go to right site -- dictionary.cambridge.org
- Go to the library -- www.lib.washington.edu
- Go for the kind of information you want -- www.npr.org

Ask, “What site provides this information?”

Search Engines

No one controls what’s published on the WWW ... it is totally decentralized.

To find out, search engines crawl Web.

- Two parts
  - Crawler visits Web pages building an index of the content (stored in a database)
  - Query processor checks user requests against the index, reports on known pages

Only a fraction of the Web’s content is crawled.

Google Advanced

Search for X

Boolean Queries

Search Engine words are independent.

- Words don’t have to occur together
- To be explicit about occurrences use Boolean queries and quotes
- Logical Operators: AND, OR, NOT
  - monet AND water AND lilies
  - “van gogh” OR gauguin
  - vermeer AND girl AND NOT pearl

Demonstration

- Google Images
- monet AND water AND lilies
- “van gogh” OR gauguin
- vermeer AND girl AND NOT pearl
Searching strategies...
- Limit by top level domains or format
- Find terms most specific to topic
- Look elsewhere for key words, e.g. bio
- Use exact phrase only when universal
- If too many hits, re-query
- "Search within results" using "." 

- Once found, ask if site is best source
- How authoritative is it?
- Can you believe it?
- How crucial is it that the information be true?
- Cancer cure for Gramma
- Hikes around Seattle

Truth on the Web
- Much Web information is wrong
- Using the Web effectively means recognizing quality information
  - Information from reliable organizations is usually preferred -- check out ownership
  - Look for accuracy, currency, ...
  - Follow links to verify that the content supports the original page

Best: Locate independent verification

A Bogus Site
http://www.burmesemountaindog.org/

A Bogus Site
http://www.MartinLutherKing.org

True Site, Bogus Implication
http://www.dhmo.org
April Fool’s Prank Site

“The US National Institutes of Health is to crack down on scientists’ brain doping” with performance-enhancing drugs such as Provigil and Ritalin, a press release declared last week. The release, brainchild of evolutionary biologist Jonathan Eisen of the University of California, Davis, turned out to be an April Fools’ prank. And the World Anti-Brain Doping Authority website that it linked to was likewise fake.”

—Nature magazine

April Fool’s Prank Site

World Anti-Brain Doping Authority (WABDA)

Welcome

• Home
• About Us
• Press
• Links
• Contact

Unit I Project

Create a bogus (fictitious) Web page

To appreciate how easy it is to fake “quality” info you will build a bogus Web page
• Modify photograph, changing its meaning
• Write misleading text
• Add “Authenticity” links, fake credentials...

Turn In Steps

• Publish your page by uploading to the Web server
• copy your Project1A files into a Project1B folder
• Submit your Word or .txt with project URL in Catalyst Collect It
• Do not touch anything in your 1A folder after the deadline
  • So the TAs can grade it
  • We will check the timestamps

Unit I Project

• Your page should look as legitimate as possible, but contain false information
• A site visitor should start out believing your site, but by the time they finish reading, they should realize that it’s a hoax
• Forget subtlety!