Chapter 5

Searching for the Truth
Searching in the Right Places

What is the right place to begin a search for information?
The Obvious and Familiar

- Where do you find tax information?
- Where do you go for directions from the University of Oregon to the Portland Airport?
- Where do you go to find the current value of Walmart stock?
Libraries Online

- University of Oregon Online Resources
  - Catalogs
  - Articles
  - Journals
  - Databases
  - Books
  - Maps & Photos
  - And more
Pre-Digital Information

• Most pre-1985 knowledge is not yet digitized.
• Most valuable information in a library is typically not online.
• Digitized documents are often missing metadata:
  - footnotes
  - references
  - appendices
“... Google's Book Search metadata are a train wreck: a mishmash wrapped in a muddle wrapped in a mess.”
"Structuring Information is as important as specifying information content ... because the value of information depends on how effectively we can use it." -- Lawrence Snyder.
Organizing Information

- Hierarchical Classification
- Trees, Roots, Branches, Leaves
- Overlapping and Partitioning of levels
- Number of levels
Searching the Web for Information

- Search Engines index the Web
  - Google
  - Bing
  - Yahoo!
  - AOL
  - AltaVista
  - Lycos
How Google Processes a Query

1. **Google User** submits a query.
2. The query is sent to **Google Web Servers**.
3. The query is then processed by the **Google Index Servers**.
4. The results are sent back to the **Google Doc Servers**.
5. The final output is displayed to the user.
How is the Web Indexed?

- Robots/Crawlers visit pages
  - Harvest all links to other pages
  - Adds new pages to TODO list
  - Creates index entries to visited page

- Crawlers miss pages
  - No link to the page
  - Dynamically created
  - Nothing readable (Just images, no HTML, etc.)
Query Processors

- Pages are indexed with keywords
- Keywords come from:
  - Document title
  - Document body text
  - Document metadata
  - Document source links
- Query words are looked up in index
Page Ranking

- Order links by relevance to user
- Counts links to the target page
  - Is the source page highly ranked?
- Checks information about the domain name
- Checks the age of the page
- Checks for duplicate copies of the page
Asking the Right Questions

- Words of Phrases?
- Logical Operators
- Narrow results by:
  - Language
  - Site or Domain
  - File Format
Logical Operators

- **AND** - The search results should contain both words around the AND operator.
- **OR** - The search results should contain either word around the OR operator.
- **NOT** - Exclude the following term/keyword.
"… Papa says, 'If you see it in THE SUN it's so.'"
– Virginia O’Hanlon (Letter to the New York Sun)
Do Not Assume Too Much

- Anyone can publish anything on the web
  - Blogs, Wikipedia, etc.
- Some of what is published is false or intensionally misleading.
- What motives might the publisher have?
- Can you trust the site?
Two Step Site Check

- Verify the company that owns domain name.
  - http://www.internic.net/whois.html
- If the whois returns another whois site, go there and repeat the process.
- You will end up getting the name and address of the person who registered the company name.
Characteristics of Legitimate Sites

- Physical Existence - Street address, phone number, email address.
- Experience - References, citations, credentials, links to related sites.
- Clarity - Well organized, easy to use.
- Currency - Updated regularly.
- Professionalism - Spelling, grammar, punctuation, and links.