


Thinking Beyond Google

Douglas J. Joubert
National Institutes of Health Library
Thursday, January 29, 2009

Outline

- How Search Engines Search
- Specialized Search Engines
- Directories
- Deep Web
- Discussion Groups
- Digital Libraries/Repositories

Search Engines



How Google Works

- Googlebot: a web crawler that finds and fetches web pages
- Indexer: sorts every word on every page and stores the resulting index of words in a huge database
- Query processor: compares your search query to the index and recommends the documents that it considers most relevant.

Google Guide (2008)

Googlebot



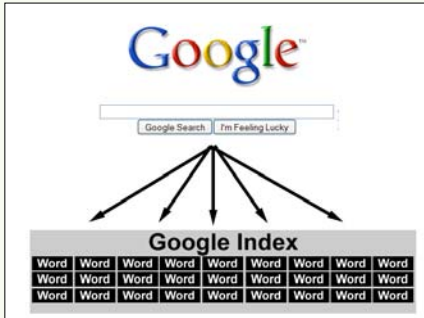
Copyright © 2003 Google Inc. Used with permission.

Google Index



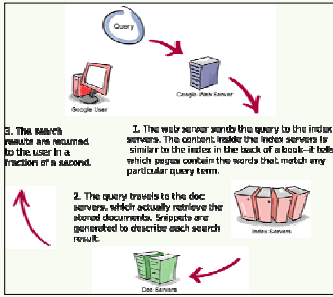
Copyright © 2003 Google Inc. Used with permission.

Google Search



Copyright © 2003 Google Inc. Used with permission.

Google Query Illustrated



Copyright © 2003 Google Inc. Used with permission.

Specialized Search Engines



Meta Search Engines

Meta Search Engine	URL
Clusty	http://clusty.com/
Dogpile*	http://www.dogpile.com/
Ithaki	http://www.ithaki.net/
Metacrawler	http://www.metacrawler.com/

Subject Search Engines

Subject Search Engine	Focus	URL
CiteSeer (x)	Computer Science	http://citeseerx.ist.psu.edu/
LawCrawler (public)	Legal	http://public.findlaw.com/
LawCrawler (prof)	Legal	http://lawcrawler.findlaw.com/

CiteSeer(x)

<http://citeseerx.ist.psu.edu/>



CiteSeer(x)



Subject (meta) Search Engines

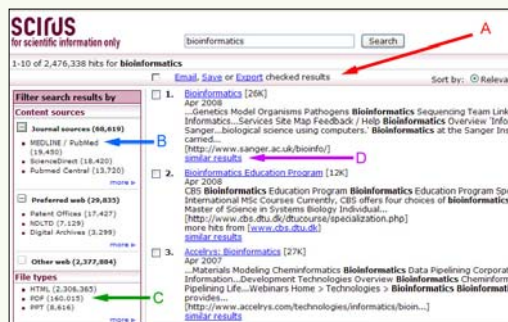
Search Engine	Focus	URL
Education Search	Education	http://www.esfs.govt.nz/dsm/esfs/search.html
Scirus	Science	http://www.scirus.com/
Science.gov	Science	http://www.science.gov/
TechXtra	Engineering & Computing	http://www.techxtra.ac.uk/
ToxSeek	Environmental Health	http://toxseek.nlm.nih.gov/

Scirus

<http://www.scirus.com/>



Scirus



Directories



Directories

- In contrast to search engines, directories list web sites by categories
- Categorization is usually based on the whole web site rather than a single page
- Generally, directories do not rank, promote or optimize sites for search engines
- Most directories are fully or partial edited by humans (SME)

How Directories Work

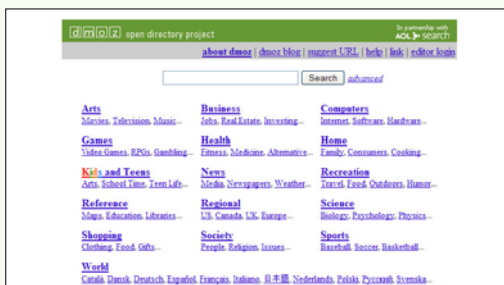
- Website owner submits a site for inclusion in the directory
- Caveat: Since submitters are choosing their own category, no true quality control
- DMOZ Example
 1. Determine whether a site is appropriate for submission to the ODP
 2. Make sure the site is not already represented in DMOZ
 3. Choose category that best describes the site

Why Use Directories

- Documents are already arranged by subject/category
- Since the websites are classified manually, most documents will be in the correct category
- Subject categories are typically arranged from broad to specific
- Resources are annotated
- Searching a smaller set of resources (quicker)

DMOZ

http://www.dmoz.org/



DMOZ - Directory Browse

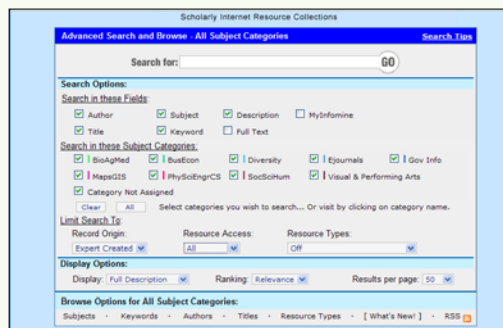


INFOMINE

http://infomine.ucr.edu/



INFOMINE - All Subject Categories



Intute

http://www.intute.ac.uk/

The screenshot shows the Intute homepage with a navigation bar at the top containing links for 'Science and Technology', 'Arts and Humanities', 'Social Sciences', and 'Health and Life Sciences'. A sidebar on the left lists various services like 'Working with Intute', 'Support materials', and 'Projects'. The main content area features an 'About Intute' section describing the service as a free online platform for education and research, and an 'Internet catalogue' section with a search bar and 'Advanced search' and 'New resources Help' links.

Intute - Categories

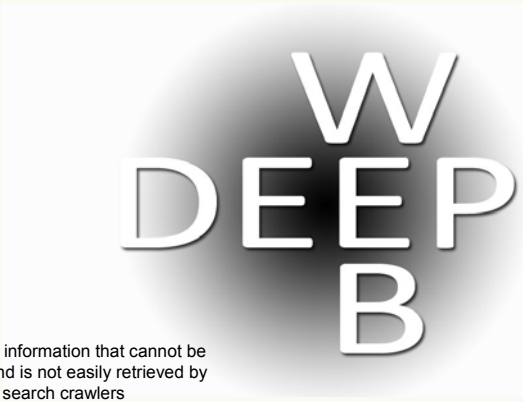
The screenshot displays the 'Categories' page with a list of subject links on the left, including 'Subject links', 'About us', 'A.Z of services', 'Internet catalogue', 'Internet training', 'Virtual Training Suite', 'Support materials', 'Intute events', and 'Additional Services'. The main content area features a search bar, a 'Browse by heading' section with a grid of subject categories (e.g., 'Biology', 'Chemistry', 'Computing', 'Earth Sciences', 'Engineering', 'Environment', 'General Sciences', 'Geography', 'Mathematics', 'Physics'), and a 'Login to MyIntute' section with an email and password field.

Intute - Sub-categories

The screenshot shows the 'Sub-categories' page with a list of subject links on the left. The main content area features a 'Browse by heading' section with a grid of sub-categories (e.g., 'Analytical Chemistry [459]', 'Inorganic Chemistry [57]', 'Organic Chemistry [773]', 'Physical Chemistry [616]', 'General Chemistry [641]', 'History of Chemistry [57]', 'Organometallic Chemistry [54]', 'Theoretical Chemistry [516]'). A search bar and 'Advanced search' and 'New resources Help' links are also present.

Intute

The screenshot displays the search results page for 'bioinformatics'. The search bar shows 'Search bioinformatics in All subjects'. The results section shows 48 results, with a list of items including 'Conference papers', 'Courses', 'FAQs', 'HE institutions and departments', 'Interactive resources', 'Journals - contents and abstracts', 'Journals - full text', 'Lecture notes', 'Working lists and discussion groups', 'Moving images', 'Other educational materials', 'Open access journals', 'Professional organisations', 'Reference sources', 'Research centres and projects', and 'Software'. Annotations A, B, and C point to the search bar, the results list, and the 'Filter by resource type' dropdown, respectively.



Describes information that cannot be indexed and is not easily retrieved by search crawlers
Dr. Jill Ellsworth (1994)

DeepWeb Documents

Document Types
Dynamically generated html pages
Non-html files
Ephemeral information
Grey Literature
Database content

Dynamically Generated Pages

- In the past, dynamically generated web pages were not found by search engines
- However, pages that used to be “invisible” are now “visible”:
 - *Pages in non-HTML formats are now converted into html*
 - *Script-based pages*
 - *Database generated pages*
 - These can be indexed if there is a stable URL somewhere that search engine crawlers can find.

<http://www.lib.berkeley.edu/TeachingLib/Guides/Internet/InvisibleWeb.html>

Non-html Formats

- Because search engine spiders are text indexers, non-text formats can be a problem
- In 2001, Google started converting and indexing pdf, Postscript, Word and Excel files

Grey Literature

- Normally defined as any publication produced outside of traditional publishing channels
A more formal definition

“open source material that usually is available through specialized channels and may not enter normal channels or systems of publication, distribution, bibliographic control, or acquisition by booksellers or subscription agents”

Grey Literature Examples

Academic papers	Dissertations
Archival Material	Government Reports
Committee Reports	Market Surveys
Conference Papers	Research Reports
Conference Proceedings	Standards
Corporate Documents	Working Reports

Selected Grey Literature Sources

GreySource Index
E-Print Network
Technical Reports and Standards
Science Accelerator
Science.gov
Virtual Technical Reports Center

GreySource Index

<http://www.greynet.org/greysourceindex.html>



E-Print Network

<http://www.osti.gov/eprints/>

NTIS

<http://www.ntis.gov/search/index.aspx>

Technical Reports and Standards

<http://www.loc.gov/rr/scitech/trs/trsover.html>

Science Accelerator

<http://www.scienceaccelerator.gov/>

Science.gov
http://www.science.gov/index.html

Virtual Technical Reports Center
http://www.lib.umd.edu/guides/techrpts.html



Digital Libraries - Repositories

- A single collection or a gateway to multiple collections
- Can include the following types of resources
 - *Digitized (i.e., scanned) books and articles*
 - *Born-digital texts*
 - *Audio files (e.g., wav, mp3)*
 - *Images (e.g., tiff, gif)*
 - *Movies (e.g., mp4, QuickTime)*
 - *Datasets (e.g., downloadable statistics files)*

OAister

<http://www.oaister.org/>



Discussion Groups

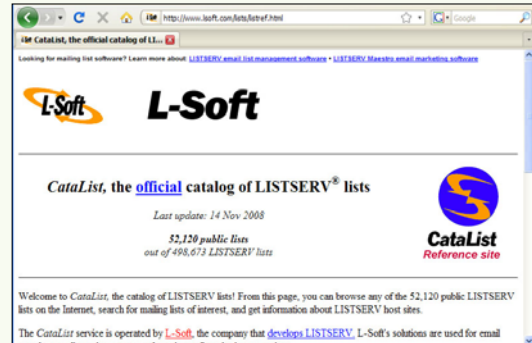


UseNet - Big 8 Categories

Domain	Category
comp.*	computer-related discussions
humanities.*	Fine arts, literature, and philosophy
misc.*	Miscellaneous topics
news.*	Discussions and announcements about news (meaning Usenet, not current events)
rec.*	Recreation and entertainment
sci*	Science related discussions
soc.*	Social discussions
talk.*	Talk about various controversial topics

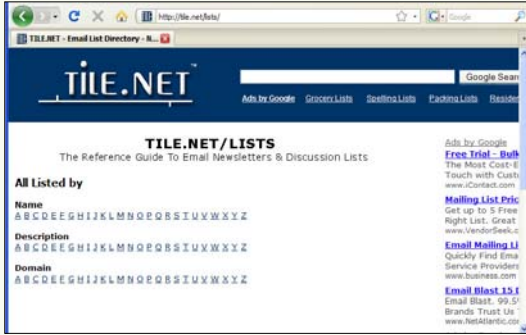
CataList

<http://www.lsoft.com/catalist.html>



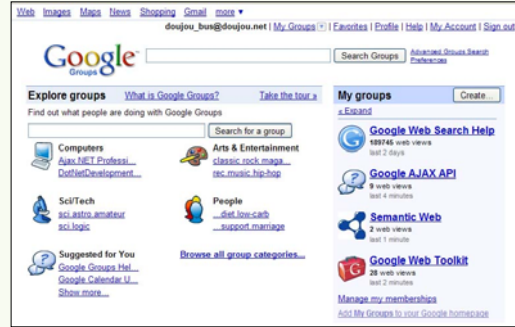
TileNet

http://tile.net/lists/



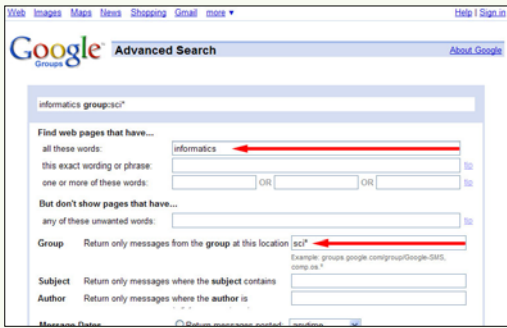
Google Groups

http://groups.google.com/

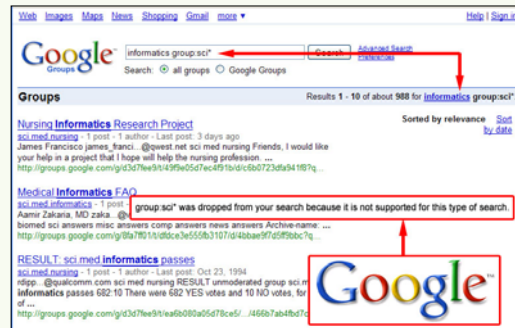


Copyright © 2003 Google Inc. Used with permission.

Google Groups - Advanced Search



Google Groups - Advanced Search



Google Groups - Results

The screenshot shows a Google Groups search result for the group 'sci.med.informatics'. Red arrows point to the following elements:

- A:** The search bar containing the text 'informatics group sci'.
- B:** The 'Subscribe to this group' link in the 'About this group' section.
- C:** The 'Options' link located above the main content area.

Blogs

Service	URL
Bloglines	http://www.bloglines.com
Feedbase	http://www.feedbase.net/
Sphere	Now enterprise search only
Syndic8	http://www.syndic8.com/
Technorati	http://technorati.com/blogs/directory/

Bloglines

<http://www.bloglines.com>

The screenshot shows the Bloglines website interface. Red arrows point to the following elements:

- A:** The search bar at the top of the page.
- B:** The 'Options' link in the left-hand navigation menu.
- C:** The 'edit subscription' link for the 'Scholarship 2.0' feed.

References

- Agosti, M. (2007). Information access through search engines and digital libraries (1st.ed. ed.). New York: Springer.
- Blachman, N., & Peek, J. (2008). Google Guide: making searching even easier. Retrieved October 08, 2008, from <http://www.googleguide.com/>
- Cabibbo, A., Grant, R. P., & Helmer-Citterich, M. (2004). The Internet for cell and molecular biologists (2nd ed.). Wymondham, Norfolk, U.K.: Horizon Bioscience.

References

- Calishain, T. (2007). Information trapping: real-time research on the web. Berkeley, Calif.: New Riders.
- Cardoso, J. (2007). The semantic web: real-world applications from industry. New York: Springer.
- Henninger, M. (2008). The hidden web: finding quality information on the net (2nd ed.). Sydney, N.S.W.: UNSW Press.

Contact Information

Douglas J. Joubert, MLIS
National Institutes of Health Library
Phone: 301-594-6282
E-mail: joubertd@mail.nih.gov
LinkedIn: <http://www.linkedin.com/in/douglasjoubert>

