

Search and the 'Net @ 2008

Trends, Challenges and Cutting-Edge Developments in Internet Search

Michael Hunter

Reference Librarian

Hobart and William Smith Colleges

For Rochester Regional Library Council

Member Libraries' Staff

Sponsored by the

Rochester Regional Library Council

Supported by Regional Bibliographic Databases and Resources Sharing (RBDB) funds granted by the

New York State Library 2007

For Today . . .

- Landscape of Search @ 2008
- Google
- Information Trapping
- Social Search
- Update on Established Services
- New Services

Web Search @ 2008

Who's crawling the Web?

- **Yahoo**
 - Owns AlltheWeb, Altavista, Inktomi, Overture
- **Google**
- **Live Search (MSN)**
- **Ask owns Teoma**
- **Gigablast**

**Worldwide Search by Region,
August 2007**

Region	Unique Searchers (000)	Searches (M)	Searches Per Searcher
Worldwide	754,459	61,036	80.9
Asia-Pacific	257,952	20,295	78.7
Europe	209,678	17,846	85.1
North America	206,278	15,976	77.4
Latin America	49,995	4,784	95.7
Middle East- Africa	30,556	2,134	69.8

Notes:

1. Searchers were aged 15 and older, using home and work locations.
2. Results exclude traffic from public computers, such as Internet cafes, and access from mobile phones and PDAs.

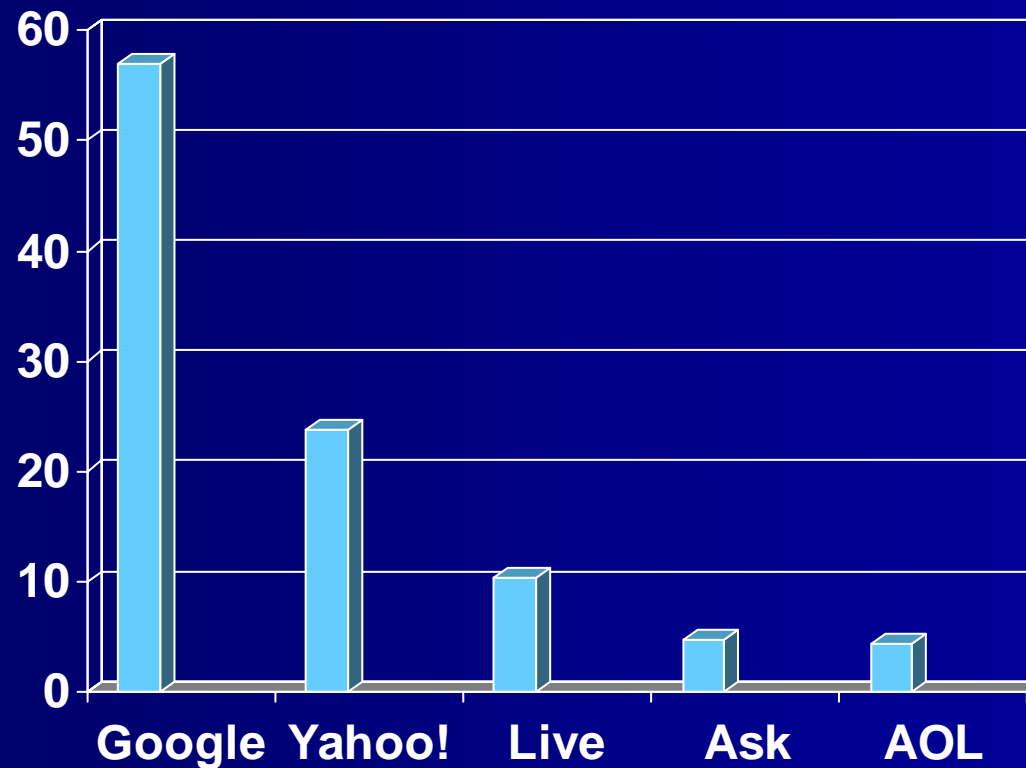
Source: comScore, 2007

**Top 10 Worldwide Search Properties,
August 2007**

Search Property	Searches (M)
Worldwide	61,036
Google	37,094
Yahoo	8,549
Baidu.com Inc.	3,253
Microsoft	2,166
NHN Corp.	2,044
eBay	1,319
Time Warner Network	1,212
Ask Network	743
Fox Interactive Media	683
Lycos Inc.	441

US Engines by Search Share

September, 2007



Size Estimates:

Google text filetypes 12/4/07

"*.html" filetype:html *in millions*

■ Html	3,120	■ Doc	3.7
■ Htm	1,170	■ Ps	2
■ Pdf	273	■ Xls	3
■ Txt	2	■ Ppt	2.5
■ Swf	3.5	■ <u>Rtf</u>	<u>2.6</u>
		TOTAL:	4,581.3

Size Estimates:

Yahoo! text filetypes 12/4/07

"*.html" + filetype limits (Adv. Search)

■ Html	2,070	■ Doc	16.3
■ Htm	611	■ RSS	227
■ Pdf	200	■ Xls	3
■ Txt	16	■ <u>Ppt</u>	<u>3</u>
■ Xml	34.3	TOTAL:	3,180.6
		Google:	4,581.3

Highlights of the Year at Google

Personalization

iGoogle and more...

- www.google.com/ig?hl=en
- Personalized portal (!!!)
- Requires a Google account
- Available in 58 countries and 26 languages
- Incorporates a user's interest profile and search behavior as search parameters

“Calculating PageRank for every person”

Disambiguation

- You’re searching for cars; You search for “jaguar”
- Results about cars appear before the animal

Location-based

- You have a default location in Google Maps
- You search for pizza or surfing lessons
- Results close to you appear first

"Pre-fab" Gadgets

- Photo gadget
- "Google-gram"
- Mini-blog
- Personal list
- YouTube video favorites
- Over 25,000 more accessible via iGoogle's Gadget Directory

Recommendations Tab

- Retrieves recommended queries, web pages, gadgets, videos, groups and news which may interest you
- Based on your personal search history and that of others with similar search histories

"Picks for You" button

- Google's "Queryless Search"
- Not tied to iGoogle, but requires an account
- Available via Google Toolbar
- Selects and displays one of 50 new sites daily, based on your search history and site popularity
- Similar to StumbleUpon, but covering more than tagging sites

Google Search by Date Range

- Google's "date" is date last visited, **not publication date**
- Blogs, News, Groups Via Advanced Search
- Web Advanced Search offers no true daterange search
- Syntax search uses Julian-style dating
daterange:startdate-enddate
- Date convertor at
<http://www.calendarhome.com/converter/>

Google Print's 2 divisions

Publisher Program and Library Project

Publisher Program

- Publishers authorize G. to scan and make searchable the full text of their books
- Users see only the page containing their search terms
- Link to purchase copy
- No download possible

Google Print's 2 divisions

Publisher Program and Library Project

Library Project (2004)

- Scan and make searchable millions of books
- For works in copyright, users see only a few sentences around search terms
- Users may browse full text of public domain works
- NOTE: Not possible to print ANY material from either Google Print project

Current Member Libraries

U. of California

Harvard

U. of Michigan

NYPL

Oxford

Stanford

U. of Virginia

U. of Wisconsin

U. of Texas (Austin)

U. Complutense (Madrid)

Princeton

Bayerische Staatsbibliothek

How Reliable is the Project?

- Predominantly pre-1923 imprints
- Quality of scanning is sometimes poor
- Editions offered are not always the best, eg. *Tristram Shandy*
- Requires huge capital investment
- Requires major effort from libraries
- Another alternative similar in scale unlikely

What full-text book search cannot provide users:

- Structured menus of research options
- Alerts of unanticipated aspects of a subject
- Retrieval of relevant items which **do not** contain the search terms
- All relevant works in foreign languages
- Reliable, systematic subject retrieval

----Thomas Mann, Lib. Of Congress

www.guild2910.org/searching.htm

How Reliable is full-text searching for books?

“Even with some of the best search and scanning technology in the world, it is unwise to ignore the bookish character of books”

-Paul Duguid

www.firstmonday.org/issues/issue12_8/duguid/index.html

Who's Scanning What: Public Domain only

- Harvard
- NYPL
- Oxford
- Princeton
- U. of Wisconsin
- U. Complutense

Who's Scanning What:

No restrictions (until courts say otherwise)

- U. of California
- U. of Michigan
- Stanford
- U. of Texas
- U. of Virginia

Copyright and the World of Books

- The set of **ALL** books that are in the public domain
- The set of **ALL** books that are copyrighted and in print (mostly in the Publisher Program)
- **ALL** other books are still in copyright, but out of print –THESE are the titles in dispute at Google's Library Project

Google and Fair Use

- “Web sites that we crawl are copyrighted. People expect their Web sites to be found, and Google searches find them. So, by scanning books, we give books the chance to be found too.” David Drummond (Google)

Publishers and the Authors Guild

- Scanning copyrighted books without permission is an infringement, even if only portions are made available to the public.
- Lawsuits filed against Google in 2005; no decision as yet

Google and “ownership” of public domain books

- “...Public domain books belong to the public and we are merely their custodians. Nevertheless, this work is expensive, so in order to keep providing this resource, we have taken steps to prevent abuse by commercial parties...”
(From the “Usage guidelines” at the point of downloading a public domain work)

Google and “ownership” of public domain books

- Use these files for personal, non-commercial use
- Refrain from automatic querying
- Maintain attribution (Google watermark)
- Keep it legal

Can Google restrict usage?

- "...We make some requests, but we are not trying to legally bind users to those requests...we are attempting to convey what we hope to be the proper netiquette for the use of these files."

Google's response to message questioning the legality of restricting public domain works

from Philipp Lenssen

<http://blogoscoped.com/archive/2007-01-10-n14.html>

WorldCat and the Project

- “Find this book in a library” link to OCLC Worldcat holdings, searchable by zip code, state, country
- Google Pilot project begun in June of 2007 to link scanned books to titles in Worldcat

Information Trapping

Information Trapping

- Automated monitoring and retrieval of information using
 - RSS feeds
 - E-mail alerts from a web site
 - Page monitors
 - Search engine alerts
- Why?
 - SAVES YOU TIME!!!

Before you set a trap ask the right questions

- Why trap for this particular topic?
 - Is this the best way to cover the topic?
- What sources do you want to search?
 - The more focused the sources, the better the results
- Is a geographic limit useful (city, state, region, etc.)?

Before you set a trap ask the right questions

- How often do you want to receive results?
 - Hourly? Daily? Weekly?
 - How much time do you have to check results?
- How do you want to receive results?
 - E-mail RSS reader
 - SMS (short message service – mobile phone or other mobile device)

Setting the trap...

- Consider your subject
- Choose the best approach and method
- Develop queries with care, test and re-evaluate them as you go
- Identify best possible resources to search
- Takes time to set up, but afterwards it runs itself

Choose the best approach

- Do the resources to be searched offer RSS?
 - YES? RSS aggregator
 - NO? E-mail alerts, page monitors
- Do you need results available via the web, or on one computer only?
 - Web based services, E-mail or SMS
 - Standalone software



RSS: What is it?

- A *broadcast* version of current content from a website, blog, news page or other source (aka "RSS Feed")
- A live, constantly updated table of contents with links to the full text, eg. a feed from NYTimes.com

How do I access RSS feeds?

- Sites with RSS feeds display a small icon (usually orange) labeled **RSS** or **XML**
- Formatted in XML (eXtensible Markup Language)
- **Aggregators** allow for reading and organizing feeds of your choosing

Weekend Edition
Sunday
Wait Wait...Don't
Tell Me!
Performance Today
World Cafe
Motley Fool Radio
More Programs

 **Hourly Newscast**
 **Program Stream**
24-Hour [schedule](#)

Contact Us
Audio Help
Discussions
NPR Everywhere
Sponsor NPR
Press Room
Submit Ideas
Corrections
Ombudsman
Site Map



Text-only

News Feeds

-  [Arts & Culture](#)
-  [Business](#)
-  [Health & Science](#)
-  [Opinion](#)
-  [People & Places](#)
-  [Politics & Society](#)
-  [U.S. News](#)
-  [World News](#)


Program Feeds

-  [All Things Considered](#)
-  [Morning Edition](#)
-  [Talk of the Nation](#)
-  [News & Notes with Ed Gordon](#)
-  [Day to Day](#)
-  [Fresh Air with Terry Gross](#)
-  [The Motley Fool Radio Show](#)
-  [Weekend Edition Saturday](#)
-  [Weekend Edition Sunday](#)
-  [World Cafe](#)

Topic Feeds

- My Yahoo!** — Web-based news aggregator; free
- NetNewsWire** — Mac OS X; commercial software, trial available
- NewsFire** — Mac OS X; commercial software, trial available
- Rocket RSS Reader** — Web-based RSS reader; free
- RSS Reader** — Windows; freeware
- SharpReader** — Windows; freeware

USING LIVE BOOKMARKS

 The **Firefox** browser supports Live Bookmarks, a system that detects RSS feeds and allows users to subscribe. Look for this icon in the bottom right corner of your browser, and click on it to subscribe. Feeds will appear in your Bookmarks list. NPR offers Live Bookmarks on all program and topic pages.



Types of RSS feeds

- Static

- Majority of feeds
- Results cannot be limited by keyword(s)
- Best for general subjects ie, Michigan football

- Keyword-based

- All results contain specified keyword(s)
- Not available for all sites
- Offered at kebberfegg.com

RSS readers/aggregators

- Browser-based
 - IE 7 "Favorites/Feeds"
 - Firefox "Live Bookmarks"
 - Safari RSS Reader
- Web-based
 - Bloglines.com Newsburst.com
 - Google Reader (www.google.com/reader)
 - My Yahoo

RSS readers/aggregators

- Standalone fee-based services

- feedreader.com newzcrawler.com
- ranchero.com/netnewswire (Mac)

NOTE: Select an aggregator that allows you to add any feed you choose, not just those available via the aggregator

Email alerts

- Service usually provided by an individual website
- Forwards new content from the site
- Some limit content to topics or keywords from your pre-set query
- Monitors entire site, not just the homepage
- Often available from news services

Finding Email alerts

- Subject metasites
- Professional organizations
- General search engines
 - “alternative energy” (e-mail alerts or email alerts)
- Yahoo Alerts
 - alerts.yahoo.com
- Google Alerts
 - www.google.com/alerts

Yahoo Alerts

alerts.yahoo.com

- Automated running of user-created saved searches
- Requires a Yahoo account
- Available for News
 - Breaking News
 - Keyword News
 - Daily News Digest

Google Alerts

www.google.com/alerts

- Available for Blogs, Web, News, Video and Groups
- Requires a profile for each alert
- Alerts also available for Google News ONLY

Page monitors

- Detect changes on specific sites
- Can return “false positives”
 - Date or visitors counters, images, weather
- Use for sites without RSS
- Use for very specific topics when RSS keyword filter not available

Web-based monitors

- WatchThatPage.com
- TrackEngine.com
- Infominder.com/webminder
Fee-based after free trial
- www.changedetect.com

Client-side monitors

- www.aignes.com/ (Windows)
- Webwatcher from chaoticsoftware.com (Mac)

Monitoring Search Engines

- Google Alerts
- Yahoo, Ask
 - Set results to maximum number
 - Create query
 - Use a page monitor to monitor changes in THAT results page
 - Obviously limited

Monitoring Search Engines

- Live.com
 - Construct query
 - Add *&format=rss* to the end
 - Subscribe as to any other rss feed

Trapping .com & .gov sites

- Many offer rss feeds, often tucked away – ebay, usagov.gov, state sites
- For sites that do not offer rss, check for e-mail alerts or use a page monitor for the “What’s New” or “Press Release” page
- Information Trapping: Real-Time Research on the Web by Tara Calishain (2006)

Social Search: The Living Web @ 2008

What is Social Search?

- “Internet wayfinding services informed by human judgment” – Chris Sherman
- Not crawler-based
- “Human judgment” – content recommended and tagged by users

Folksonomies and more...

User-created metadata

Tags

Comments

Reviews

Feedback

Ratings

- YahooShopping, eBay – “Feedback”
- Blogs, Wikis, Social networking sites

Social Networking Sites Vertical Search Engines +

- Discovery of and sharing online resources
 - Del.icio.us
- BtoB
 - LinkedIn.com
- Travel and hospitality
 - TripAdvisor.com
- "Community"
 - Facebook, MySpace
 - YouTube, Digg

Social Browsers

Flock – A Social Web Portal

- www.flock.com
- Access to all major social networking sites
- Manages photos, bookmarks, blogs and tagging tools from all services in **one** system
- Facilitates sharing across many networking sites

Social Web Homepage

StumbleUpon

- www.stumbleupon.com
- Focuses on site discovery and tagging
- Personal homepage format
- Controls reside on your browser as buttons
- “StumbleUpon” button takes you to sites based on a profile of interests you create
- “Learns” your preferences, refining future sites you are taken to
- Searches user-generated tags ONLY

Meta for Tagging Sites: thagoo.com

- Searches Furl, BlueDot, del.icio.us, BlogMarks, Blinklist, MyWeb Yahoo, MisterWong and more...
- Keyword search
- Sort results by most recent or most popular
- RSS available

Google and Social Search

Google Coop

- www.google.com/coop/topics/?show=all
- Users invited to contribute sites and categorization
- Topics now include Health, Travel, Cars, Gaming, Photography and A-V
- Inclusion here MAY impact a site's ranking in regular searches

Health labels

Select a category:

Overview	Causes/risk factors	From medical authorities	Children
Treatment	Prevention	For health professionals	Teens
Symptoms	Research overview	For patients	Women
Tests/diagnosis	Practice guidelines	Support groups	Men
Complications	Patient handouts	Alternative medicine	Seniors
Prognosis	Clinical trials	Continuing education	

Try it out:

Search

Contributors

1 - 8 of 26 [N](#)

[Mayo Clinic](#)



[Stanford Hospital and Clinics](#) [New England Journal of Medicine](#)



[New England Journal of Medicine](#)



[Kaiser Permanente](#)



The Latest at Established Services

New Services

Yahoo's Search Assist

- Ajax-based service that “suggests” terms as you type
- Activate by clicking blue arrow below the search box before searching
- Also offers “Explore Concepts”, phrases highly associated with the search terms

Yahoo's MyWeb and Local

- More active registered users than Google
- Personal data greater than Google
- Acquired del.icio.us in 2005
- MyWeb offers tagging and sharing bookmarks with others
- Local search relies on IP address and personal data (if available)

Yahoo, media and the living web

- Web page alerts
- Feed/Blog alerts
- Answers (human service)
- Podcasts
- Y!Q (contextual search)

Big post-Kyoto meeting set for Montreal

Last Updated Thu, 17 Feb 2005 00:11:27 EST

CBC News

MONTREAL - Montreal will play host to about 10,000 scientists, environmentalists and politicians from 180 countries around the world this fall, when a United Nations conference on climate change takes place in the city.

- INDEPTH: The Kyoto Protocol



Results automatically refined using the following terms

from: <http://www.cbc.ca/sto...o-montreal050216.html>

[montreal montreal](#) [united nations conference](#) [montreal](#) [climate change](#)

SPONSOR RESULTS

- [Chevron Energy Debate: Climate Change](#) Get info on environment and energy issues such as **climate change** and how they effect the world's energy demands. Chevron invites you to join an online forum about the energy issues we face.
www.willyoujoinus.com
- [Prevent Climate Change with MyClimate](#) Fly Green Class with Myclimate – Calculate your greenhouse gas emissions, invest in certified **climate**-friendly projects, and neutralize the negative impacts associated with your air travel.
www.my-climate.com

 [Related Yahoo! Images for montreal montreal](#)



1. [Montreal Conference December Teachers](#) 

... Canada. Formal Conference - Teaching. December 2000 ... Montreal - Montreal to host UN climate change conference - MONTREAL " Montreal will host a major United Nations conference ...

www.educators-nfo.com/montreal-conference-december-teachers.html - 10k - [Cached](#) - [More from this site](#) - [More like this](#)

2. [Major Global Warming Conference Begins in Montreal](#) 

... On the first day of a United Nations conference on the crisis of global warming, Canadian ... California are creating...Major Global Warming Conference Begins in Montreal MONTREAL, Nov ...

environment.about.com/b/a/222880.htm - 25k - [Cached](#) - [More from this site](#) - [More like this](#)

3. [Montreal - Wikipedia, the free encyclopedia](#)  - [Translate this page](#)

... 2005, Montreal hosted the United Nations Climate Change Conference, the first meeting joint ... located in Longueuil, southeast of Montreal. Montreal also hosts the headquarters of ...

en.wikipedia.org/wiki/Montreal - [More from this site](#) - [More like this](#)

4. [German Consulate General Montreal - Montreal to host next global climate change conference](#) 

Navigation ... held in conjunction with the 11th Conference of the ... United Nations



MSN's Live.com

- Database increasing
- "Rich Answers"
- Image search enhancements
filter:face filter:portrait filter:bw
- NLP question processing improved

New Services

- ChaCha.com

- Live “guides” take your query and return customized results
- Requires account (free)

- TVEyes.com

- Searches full text of every word from major TV news broadcasts
- Errors exist

Fox News

CBS

MSNBC

CNN

Reuters

BBC

ChaCha^{BETA}[®]

search + brainpower

web

news

images

video

audio

ChaCha Search

Search with Guide

[Why ChaCha?](#) ⓘ | [Toolbar Download](#) | [Register to use a Guide](#)

Searchcrystal.com: Visual Metasearch

- Options include List View, Spiral View and Cluster Display
- Results common to more engines appear in the center
- Color=source engine
- Shape=number of engines retrieving the page
- Size=rank position
- Web, Blogs, Images, Tagging sites and more...

Advancing Technologies and Search

Tafiti: live.com's experimental interface

- For research that needs multiple queries and sessions
- Offers visualization, tagging, storage and sharing of research
- Live.com ID required for saving and sharing
- Requires Silverlight 1.0 RC presentation software (free download)



sign in | faq | terms of use | privacy | feedback



Web

1-30 of 7,760,000 results



filter these results

SEARCH finger lakes

GO

NY Tourism & T...

www.fingerlakes.org

The official site of the region. Includes events, attractions, restaurants, accommodations and wineries.

New York Tourism - ...

www.visitfingerlakes.com

New York Tourism - Official tourism information web site for Finger Lakes Region's Lake Country; information on accommodations, attractions, recreation, restaurants, wineries

Finger Lakes Commu...

www.fingerlakes.edu

Main campus is on Lake Shore Drive in Canandaigua, with extension centers in Geneva, Newark, and Victor, NY.

Future Students : Fin...

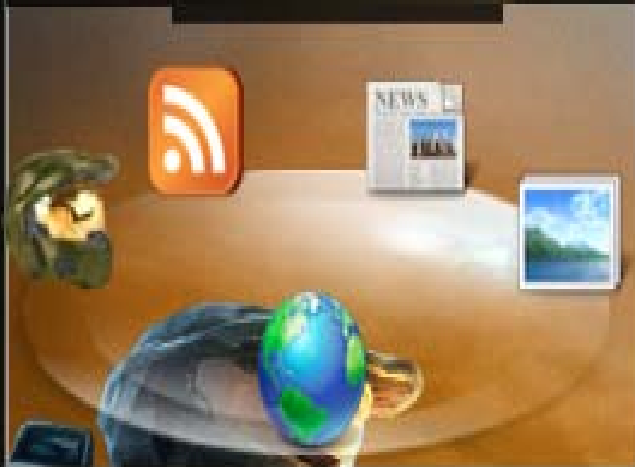
www.fingerlakes.edu

Thinking about community college? Find out what makes Finger Lakes Community College the Right Place to Start ... outstanding academic programs, comprehensive support services and ...

Thoroughbred Horse...

www.fingerlakesrace.com

Horse racing schedules, statistics, promotional events, downloadable coupons, wagering instructions.





NY Tourism & Travel - Finger Lakes
New York Tourism - Finger Lakes Vi...

New York Horse Racing - Finger Lak...
Thoroughbred Horse Racing Trac...
New York Tourism - Finger Lakes Vi...
Finger Lakes Runners Club - Upstat...
Friends of the Finger Lakes

Finger Lakes Realty - Finger Lakes R...
New York Tourism - Visit Coming a...

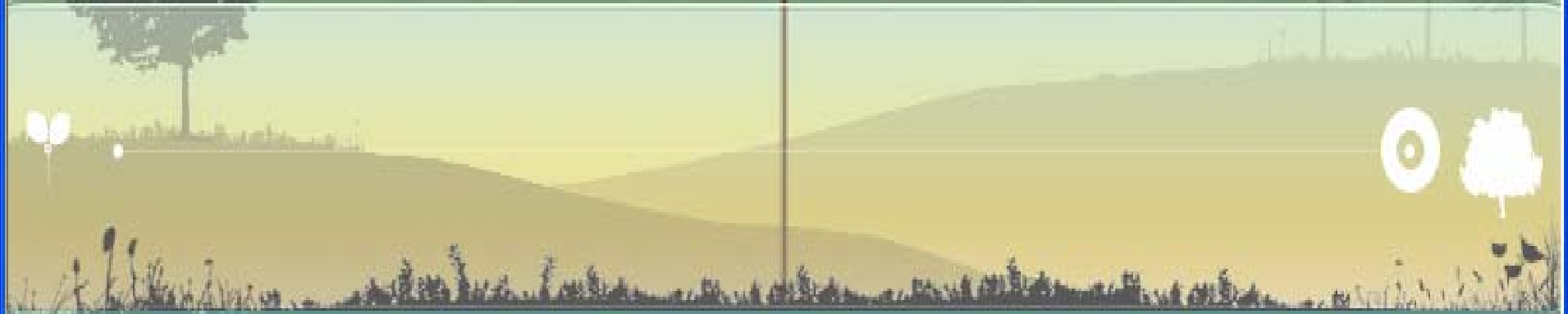
Finger Lakes Culinary Bounty
Finger Lakes Trail Conference
Finger Lakes Hotels
Finger Lakes - Wikipedia, the free e...

Finger Lakes Lodging - Skaneateles ...
Finger Lakes Winery Tours

Finger Lakes Limousine

Fingerlakes.net

Finger Lakes Wine, Finger Lakes Win...



AJAX

Asynchronous JavaScript And Xml

- Original Web – static documents in html with linkage capability
- Dynamic Web – “Active” documents created as requested from databases with time-specific data and/or user-specific needs.
- CGI, JavaScript, asp, asp.net, php

AJAX

Asynchronous JavaScript And Xml

- Used for interactive web applications
- Small parts of a page can be refreshed without a complete reload
- "As if your were searching the database, but transparently"
- "Sliders" and auto-populated forms
- Search engines cannot crawl content in Java

Wikiasari:

Quick rummaging search

- "User ranked results"
- Open source SE by Jimmy Wales and Amazon
- Initial results ordered with algorithms a la Google
- Users reorder results, which will be used in ranking of future similar searches

Wikiasari:

Quick rummaging search

- Strength is in general search topics
- Deep, complex or unusual searches will not benefit as much
- Intended to rival Google and Yahoo
- Not yet launched
- Search.wikia.com/wiki/Search_Wikia

Semantic Search Systems

- Understand the user's query
- Understand Web text
- Bring these together for query results that are contextually relevant
- Algorithms that match the meanings and not just the words
- Natural Language Processing
- Concept Mapping

Semantic Search Systems

- Expensive and time-consuming for general web search; more possible in subject-specific contexts

“What is palladium used for?”

- Link-based crawler results:
 - London’s Palladium Theatre
- Include the concept map “used for”
 - Sites about the element palladium
- Hakia, Powerset, Cognition Search

Post Search

What do we do AFTER a Google search?

- The search engine size wars are over
- WANTED: Services that help manage, share and update
 - Web search results
 - Trapped information
 - Tagged sites WITH
 - scalability* *confidentiality* *“collaborability”*
 - across all platforms, devices and file formats*

Thank You!

Michael Hunter
Reference Librarian
Hobart and William Smith Colleges
Geneva, NY 14456

(315) 781-3552

hunter@hws.edu