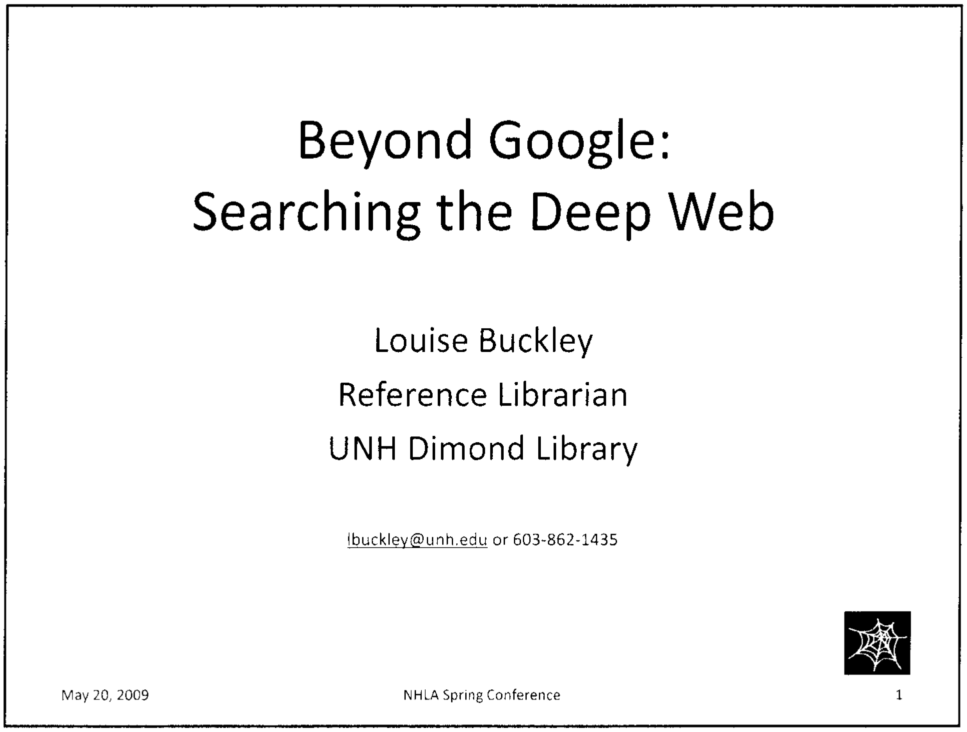
5-20-2009

"Beyond Google: searching the deep web."

Louise Buckley

University of New Hampshire, lousie.buckley@unh.edu



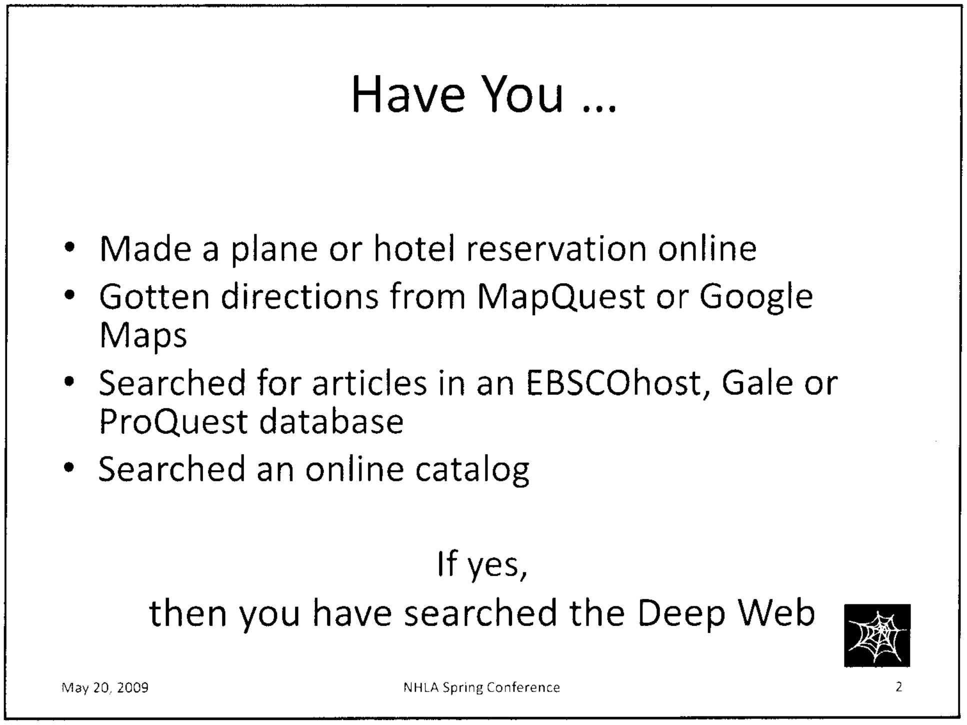
Beyond Google: Searching the Deep Web

Louise Buckley

Reference Librarian

UNH Dimond Library

lbuckley@unh.edu or 603-862-1435



Have You …

- Made a plane or hotel reservation online
- Gotten directions from MapQuest or Google Maps
- Searched for articles in an EBSCOhost, Gale or ProQuest database
- Searched an online catalog

If yes,
then you have searched the Deep Web

Visible Web = Surface Web = Shallow Web

Findable by general purpose search engines

Deep Web = Invisible Web = Hidden Web = Dark Web = Cloaked Web

Not findable by general purpose search engines

Timeline of Events Related to the Deep Web

http://papergirls.wordpress.com/2008/10/07/timeline-deep-web/

Deep Web larger than Surface Web

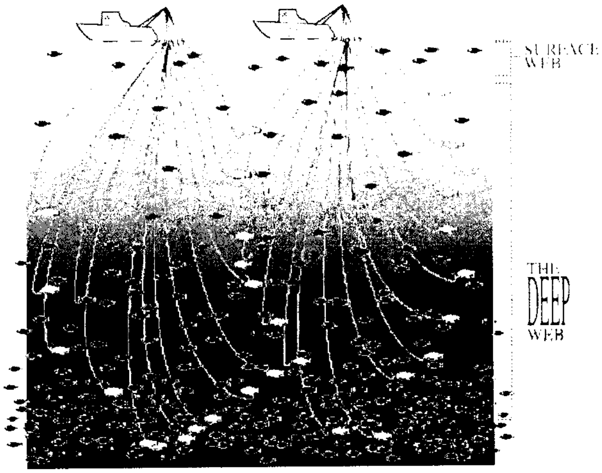
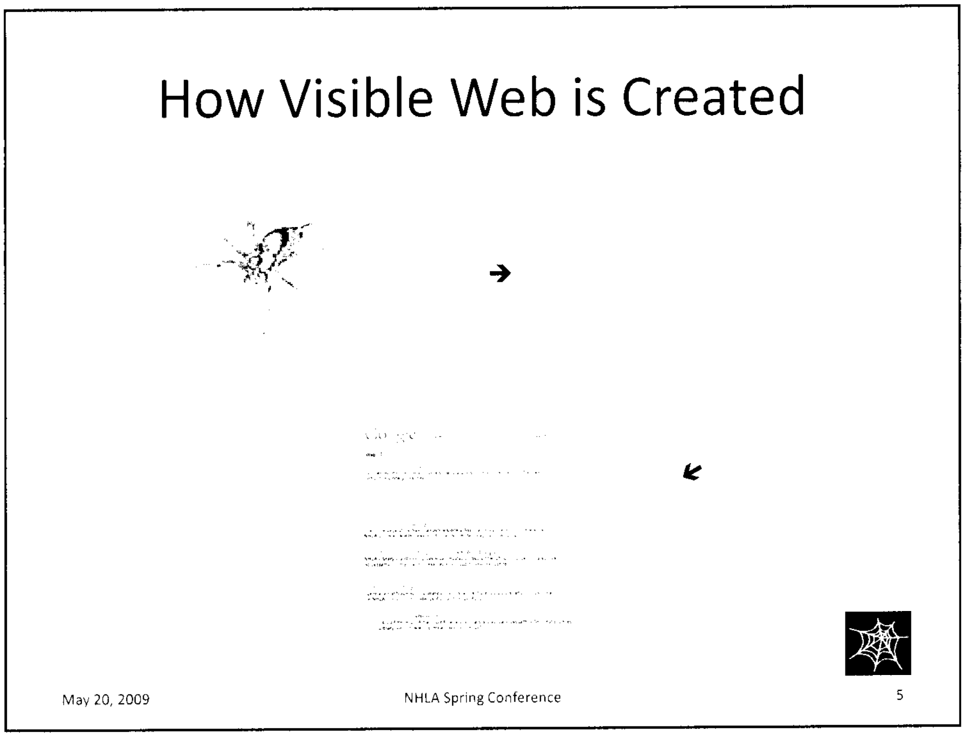


Image Source: http://www.mkbergman.com/?p=458

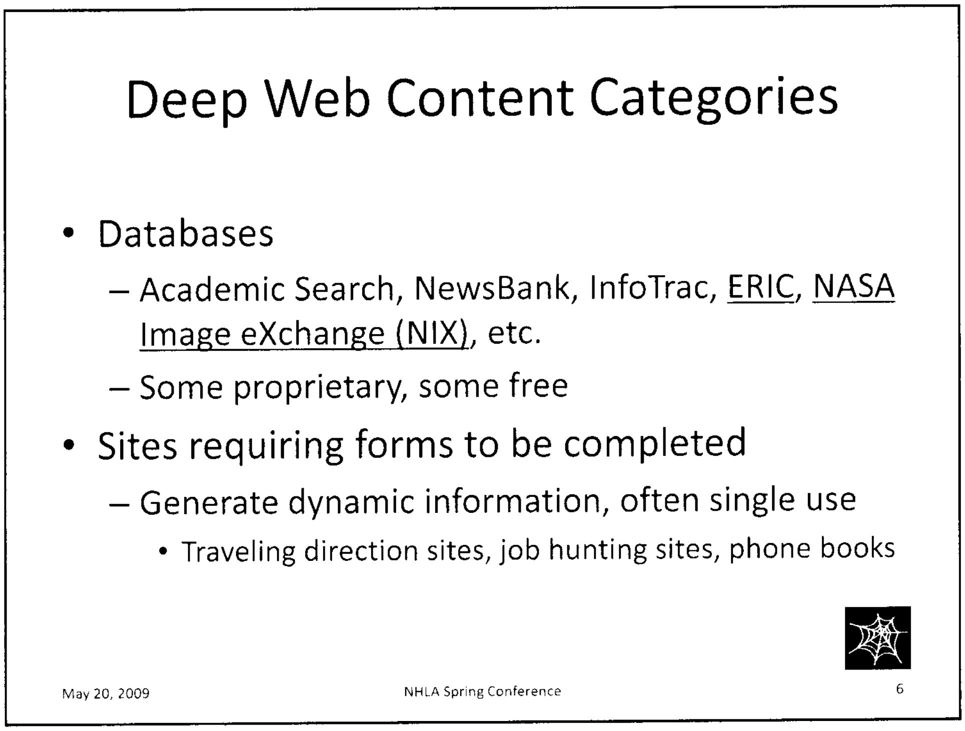
2007 Estimates:

http://netforbeginners.about.com/cs/secondaryweb1/a/secondaryweb.htm



How Visible Web is Created

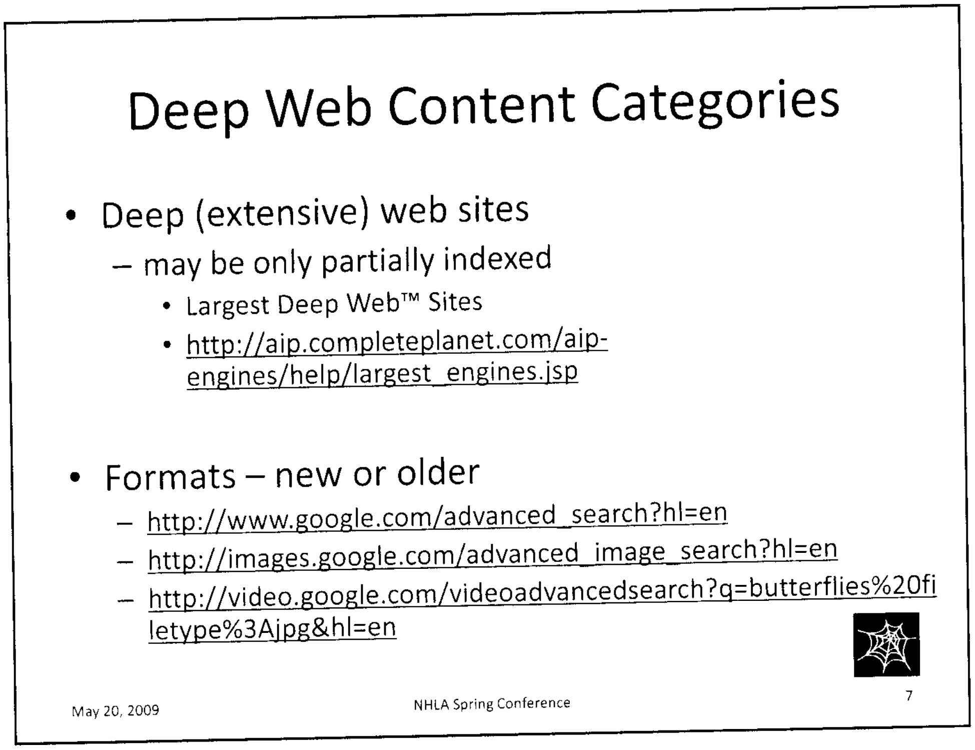
May 20, 2009 NHLA Spring Conference 5



Deep Web Content Categories

- Databases
  - Academic Search, NewsBank, InfoTrac, ERIC, NASA Image eXchange (NIX), etc.
  - Some proprietary, some free
- Sites requiring forms to be completed
  - Generate dynamic information, often single use
    - Traveling direction sites, job hunting sites, phone books

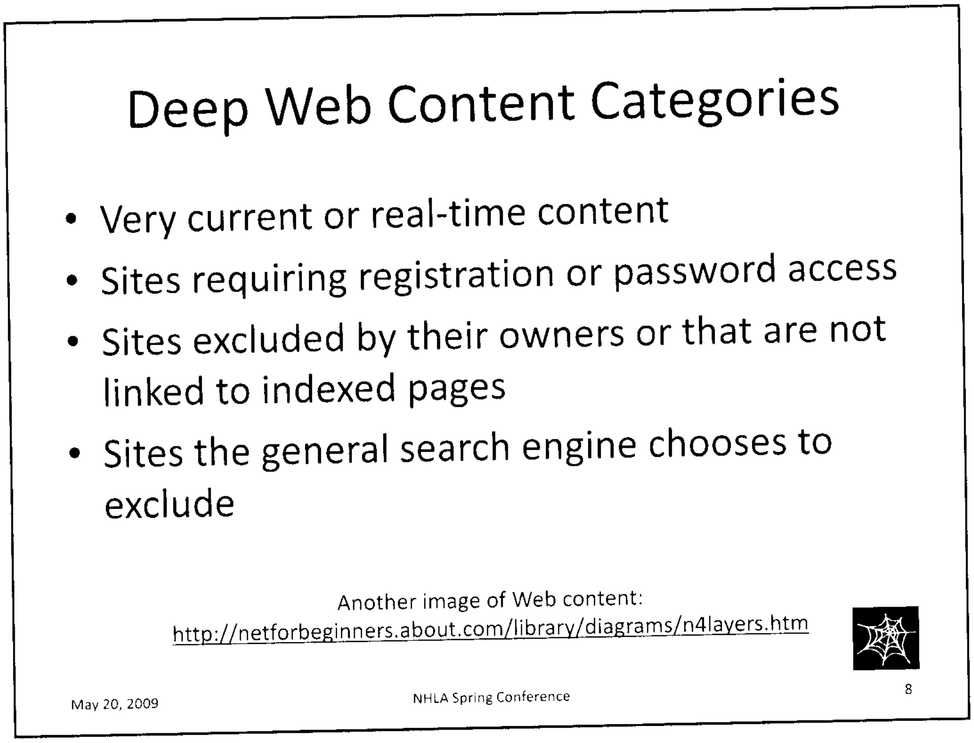
May 20, 2009 NHLA Spring Conference 6



Deep Web Content Categories

- Deep (extensive) web sites
  - may be only partially indexed
    - Largest Deep Web™ Sites
    - http://aip.completeplanet.com/aip-engines/help/largest_engines.jsp
- Formats – new or older
  - http://www.google.com/advanced_search?hl=en
  - http://images.google.com/advanced_image_search?hl=en
  - http://video.google.com/videoadvancedsearch?q=butterflies%20filetype%3Ajpg&hl=en

May 20, 2009 NHLA Spring Conference 7

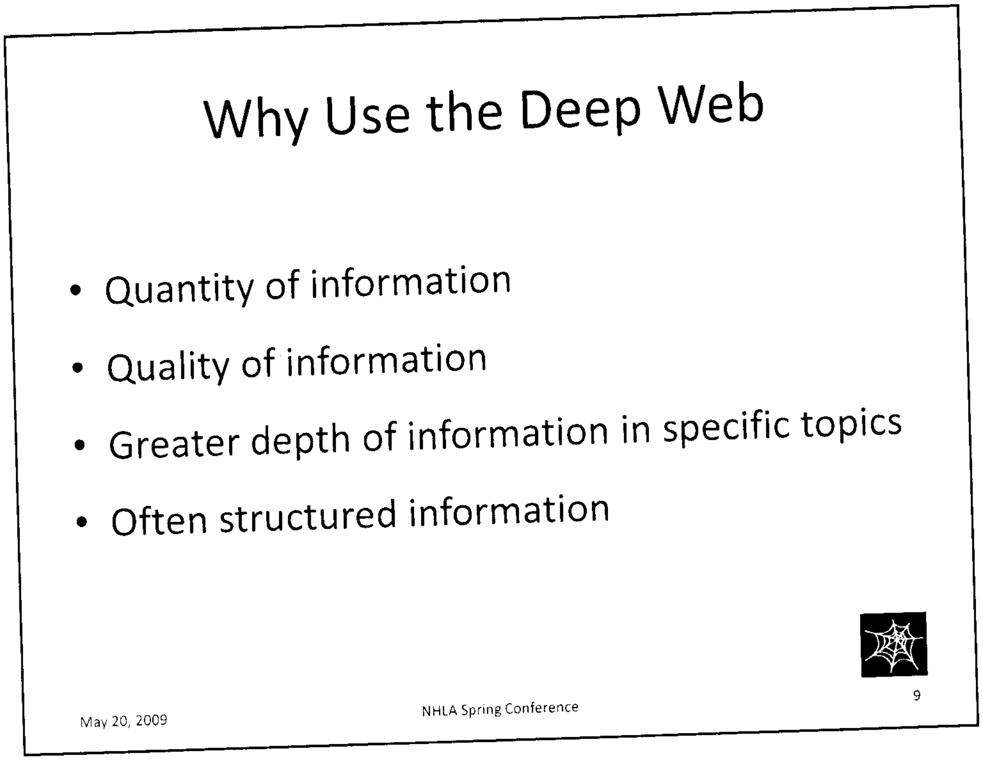


Deep Web Content Categories

- Very current or real-time content
- Sites requiring registration or password access
- Sites excluded by their owners or that are not linked to indexed pages
- Sites the general search engine chooses to exclude

Another image of Web content:
http://netforbeginners.about.com/library/diagrams/n4layers.htm

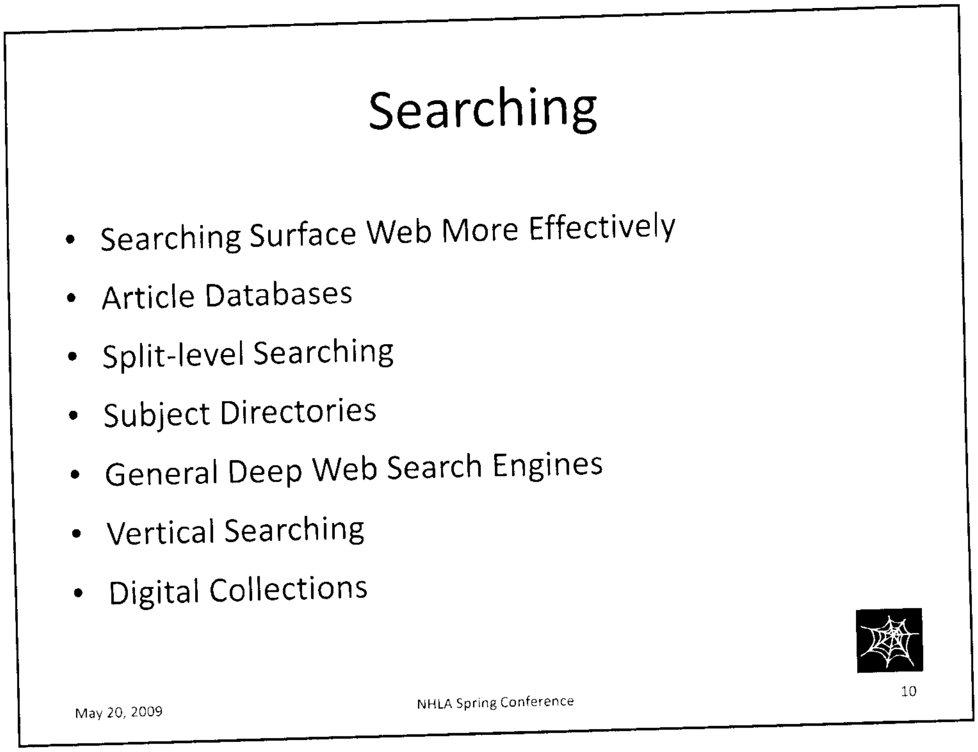
May 20, 2009 NHLA Spring Conference 8



Why Use the Deep Web

- Quantity of information
- Quality of information
- Greater depth of information in specific topics
- Often structured information

May 20, 2009 NHLA Spring Conference 9



Searching

- Searching Surface Web More Effectively
- Article Databases
- Split-level Searching
- Subject Directories
- General Deep Web Search Engines
- Vertical Searching
- Digital Collections

May 20, 2009 NHLA Spring Conference 10

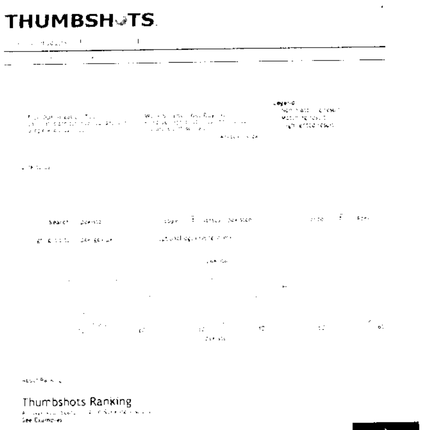
More Effective Surface Web Searching

ranking.thumbshots.com

Try more than one general purpose search engine

- Learn strengths/weaknesses/coverage
- "Different Engines, Different Results" http://www.infospaceinc.com/onlineprod/Overlap-DifferentEnginesDifferentResults.pdf

[Article on Thumbshots ranking tool]



THUMBSH◌TS

Thumbshots Ranking

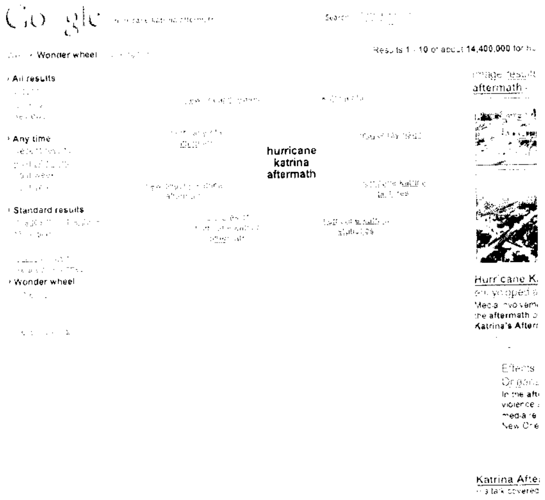
More Effective Surface Web Searching

- Use advanced search features
  - Domain/site restriction - .edu or .gov or .org
  - Other refinements – file format, page date, where keywords appear, language, usage rights
  - Advanced search features available for image, news, video searches
    - http://www.google.com/advanced_search?hl=en
    - http://search.yahoo.com/

Searching the Surface Web More Effectively

Explore "Show Options" on Google results page:

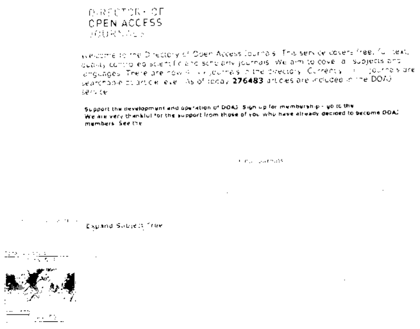
- › All results
  - Videos
  - Forums
  - Reviews
- › Any time
  - Recent results
  - Past 24 hours
  - Past week
  - Past year
- › Standard results
  - Images from the page
  - More text
- › Standard view
  - Related searches
  - Wonder Wheel
  - Timeline



Sample Search: hurricane katrina

Article Databases

- Subscription databases
  - EBSCOhost, Newsbank, etc.
- Google Scholar
- Free databases
  - Directory of Open Access Journals (DOAJ) - http://www.doaj.org/
  - HighWire Press - http://highwire.stanford.edu/lists/freeart.dtl
  - PubMed Central - http://www.pubmedcentral.nih.gov/


DIRECTORY OF OPEN ACCESS JOURNALS

Welcome to the Directory of Open Access Journals. This service covers free, full text, quality controlled scientific and scholarly journals. We aim to cover all subjects and languages. There are now [illegible] journals in the directory. Currently [illegible] journals are searchable at article level. As of today 276483 articles are included in the DOAJ service.

Support the development and operation of DOAJ. Sign up for membership - go to the

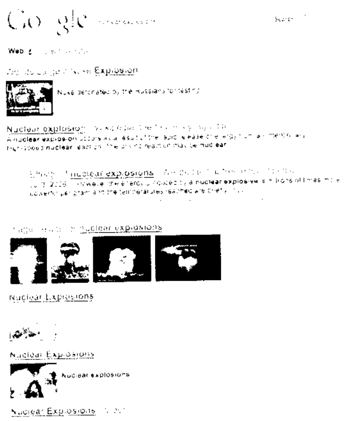
We are very thankful for the support from those of you who have already decided to become DOAJ members. See the

Split-Level Searching

- First Level
  - Use general purpose search engines to locate directories, specialized databases or search engines
- Second Level
  - Search the directory, database, search engine, etc.

First level search strategy:
include "database" with topic term(s) in search

nuclear explosions



Google

Web

Explosion

Nuke detonated by the Russians for testing

Nuclear explosion

A nuclear explosion occurs as a result of the rapid release of energy from an intentionally high-speed nuclear reaction. The driving reaction may be nuclear

nuclear explosions

nuclear explosions

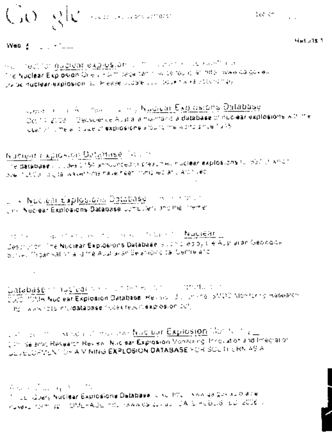
Nuclear Explosions

Nuclear Explosions

Nuclear explosions

Nuclear Explosions

nuclear explosions database



Google

Web

Results

nuclear explosion

The Nuclear Explosion

nuclear explosion

Nuclear Explosions Database

Geoscience Australia maintains a database of nuclear explosions with the location, time and size of explosions around the world since 1945

Nuclear Explosion Database

The database includes 2150 announced or presumed nuclear explosions

Nuclear Explosions Database

Nuclear Explosions Database

Nuclear

Description: The Nuclear Explosions Database is compiled by the Australian Geological Survey Organisation and the Australian Seismological Centre and

Database nuclear

Nuclear Explosion Database

database nuclear explosion

Nuclear Explosion

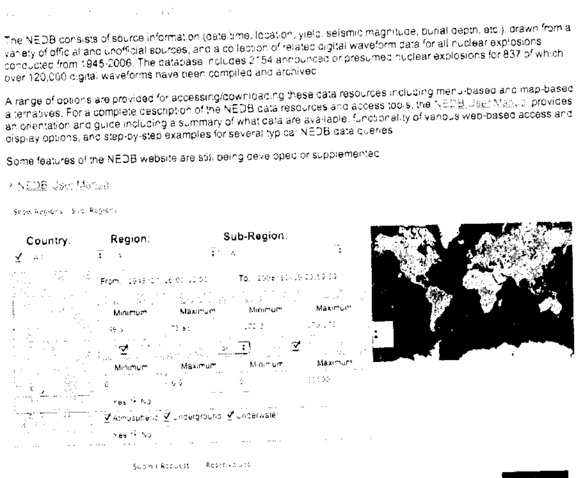
Seismic Research Review: Nuclear Explosion Monitoring: Innovation and Integration

DEVELOPMENT OF A MINING EXPLOSION DATABASE FOR SOUTHERN ASIA

Nuclear Explosions Database

Second level: search the database or portal

Nuclear Explosion DataBase (NEDB)
http://www.rdss.info/database/nedb/nedb_ent.html



The NEDB consists of source information (date/time, location, yield, seismic magnitude, burial depth, etc.), drawn from a variety of official and unofficial sources, and a collection of related digital waveform data for all nuclear explosions conducted from 1945-2006. The database includes 2154 announced or presumed nuclear explosions for 837 of which over 120,000 digital waveforms have been compiled and archived

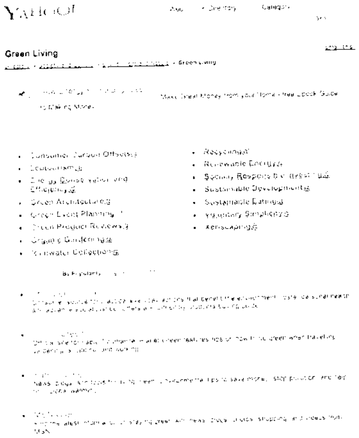
A range of options are provided for accessing/downloading these data resources including menu-based and map-based alternatives. For a complete description of the NEDB data resources and access tools, the NEDB User Manual provides an orientation and guide including a summary of what data are available, functionality of various web-based access and display options, and step-by-step examples for several typical NEDB data queries

Some features of the NEDB website are still being developed or supplemented

May 20, 2009 NHLA Spring Conference 17

Subject Directories

- About.com
  - http://www.about.com/
- Yahoo Directory
  - http://dir.yahoo.com/

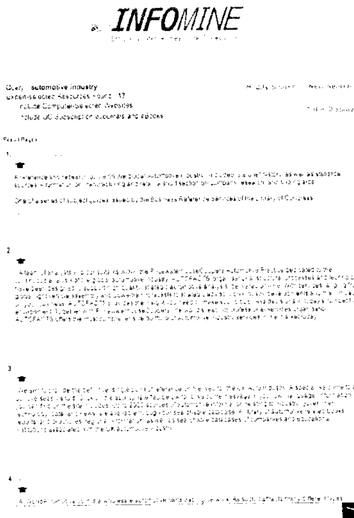


- Consumer Carbon Offsets
- Ecotourism
- Energy Conservation and Efficiency
- Green Architecture
- Green Event Planning
- Green Product Reviews
- Organic Gardening
- Rainwater Collection
- Recycling
- Renewable Energy
- Socially Responsible Investing
- Sustainable Development
- Sustainable Eating
- Voluntary Simplicity
- Xeriscaping

May 20, 2009 NHLA Spring Conference 18

Subject Directories

- Librarians' Internet Index
  - http://www.lii.org
- INFOMINE
  - http://infomine.ucr.edu/
- Intute
  - http://www.intute.ac.uk/



INFOMINE

Query: automotive industry
Expert-selected Resources Found: 37
Include Computer-selected Websites
Include UC Subscription eJournals and eBooks

1.

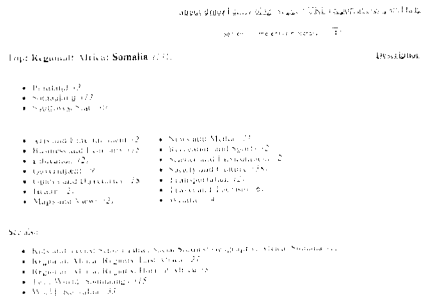
2

3

4

Subject Directories

- Open Directory Project (DMOZ)
  - http://www.dmoz.org
- World Wide Web Virtual Library
  - http://vlib.org/



Top: Regional: Africa: Somalia

Description

- Arts and Entertainment
- Business and Economy
- Education
- Government
- Guides and Directories
- Health
- Maps and Views
- News and Media
- Recreation and Sports
- Science and Environment
- Society and Culture
- Transportation
- Travel and Tourism
- Weather

See also:

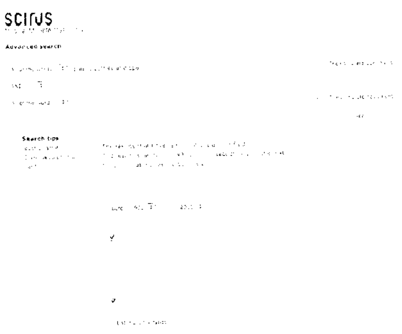
General Deep Web Search Engines

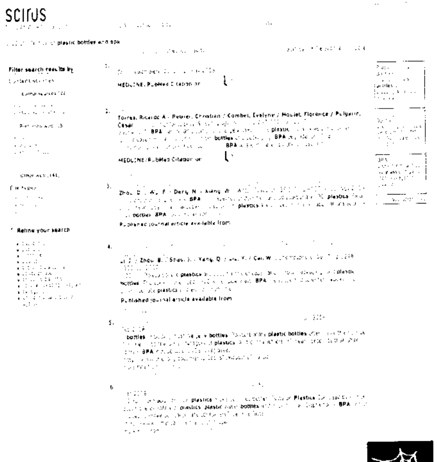
- CompletePlanet (70,000+ sites last harvested in 2004)
  http://www.completeplanet.com/
- Incy Wincy (content appears similar to DMOZ)
  - http://www.incywincy.com/

May 20, 2009 NHLA Spring Conference 21

Vertical Searching

Scirus - http://www.scirus.com/



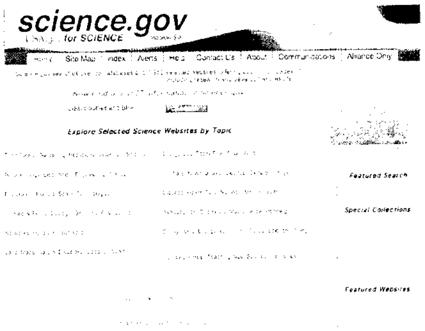


May 20, 2009 NHLA Spring Conference 22

Vertical Searching

http://science.gov/

http://scienceresearch.com/



science.gov

for SCIENCE

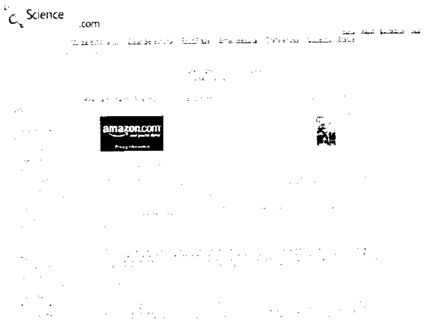
Home | Site Map | Index | Alerts | Help | Contact Us | About | Communications | Alliance Only

Explore Selected Science Websites by Topic

Featured Search

Special Collections

Featured Websites



Science .com

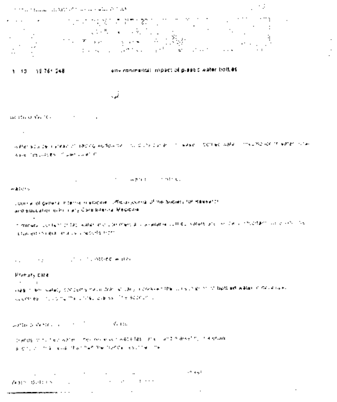
amazon.com

May 20, 2009 NHLA Spring Conference 23

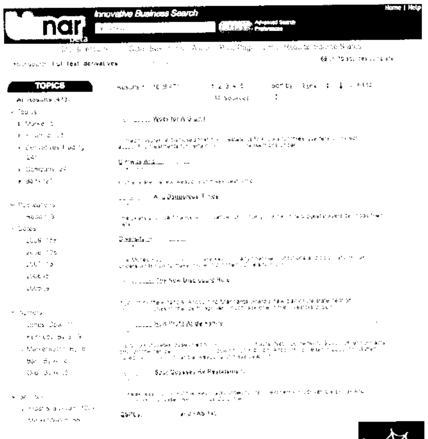
Vertical Searching

deepdyve.com

biznar.com/



environmental impact of plastic water bottles



nar

Innovative Business Search

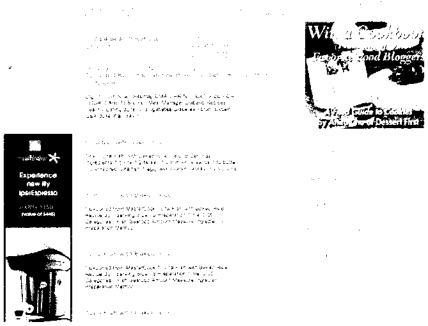
Home | Help

TOPICS

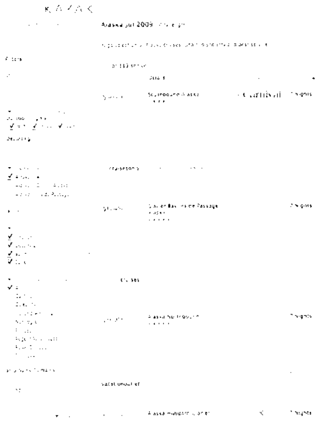
May 20, 2009 NHLA Spring Conference 24

Vertical Searching

Foodieview.com – a recipe search engine



Kayak.com – a travel search engine


KAYAK

Alaska Jul 2009

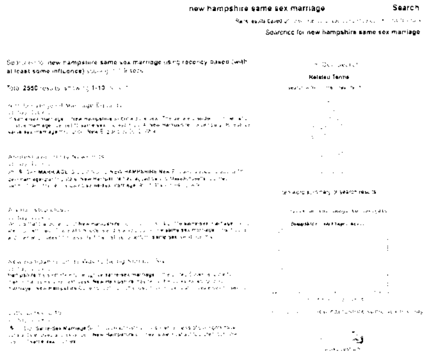
Filters

Southbound Alaska

Glacier Bay Inside Passage

Vertical Searching

http://www.blogscope.net/


new hampshire same sex marriage Search

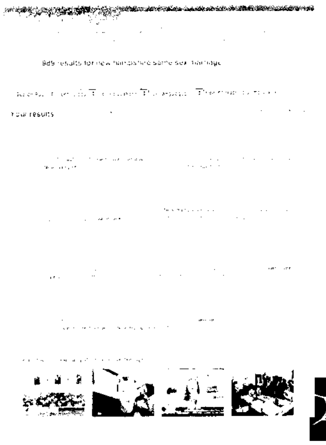
Searched for new hampshire same sex marriage

new hampshire same sex marriage using recency based (with at least some influence)

Total 2550 results showing 1-10

Related Terms

http://technorati.com/


Your results

Vertical Searching

- Music – not just iTunes
  - allmusic.com
  - Songza.com
  - iLike.com
- Videos – not just YouTube
  - blinkx.com
  - hulu.com
- Additional Suggestions from Session Attendees:
  - search.twitter.com
  - Pandora.com (Internet radio service based on The Music Genome Project)

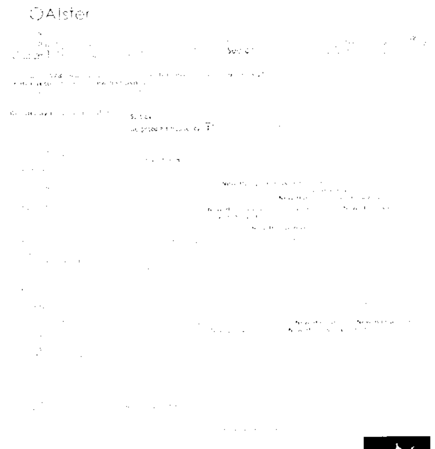
Digital Collections

- May contain
  - 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)
- May be freely accessible, restricted, or a combination
- A Compendium of Digital Collections *(A project of the UNH Library Digital Collections Initiative)* - http://digitalcollections.wordpress.com/
- Library of Congress Digital Collections & Services - http://www.loc.gov/library/libarch-digital.html
- Collaborative Digitization Programs in the United States – http://frank.mtsu.edu/~kmiddlet/stateportals.html
- PolicyArchive.org

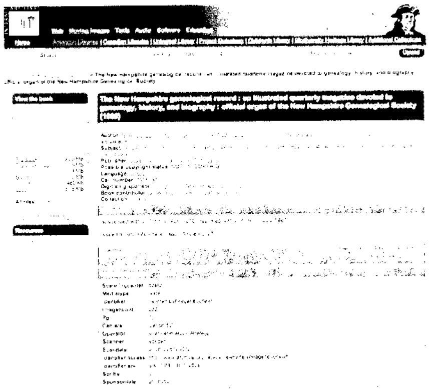
Digital Collections

- OAI – Open Archives Initiative
- OAIster - http://www.oaister.org/
  - Union catalog of digital resources
  - 20,928,590 records from 1,112 contributors
- Goals
  - One-stop "shopping" for academically-oriented digital resources
  - Access to digital materials in addition to searching of records for these resources

OAIster Search



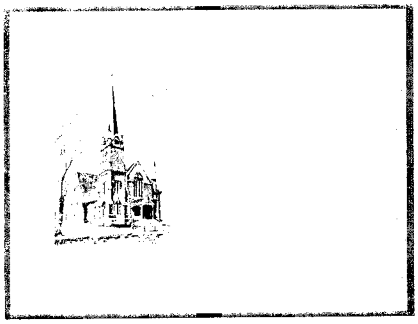
OAIster Search



The New Hampshire genealogical record : an illustrated quarterly magazine devoted to genealogy, history, and biography : official organ of the New Hampshire Genealogical Society

Author
Volume
Subject
Publisher
Possible copyright status
Language
Call number
Digitizing sponsor
Book contributor
Collection

Scanningcenter
Mediatype
Camera
Operator
Scanner
Scandate
Identifier-access
Identifier-ark
Scribe
Sponsordate



Internet Archive

- Wayback Machine – http://www.archive.org/web/web.php
- Open Library Project
- Open Content Alliance – digitization of out-of-copyright books and accessible through a permanent archive
  - Text archive - http://www.archive.org/details/texts

Internet Archive

Sample text search:
durham nh report

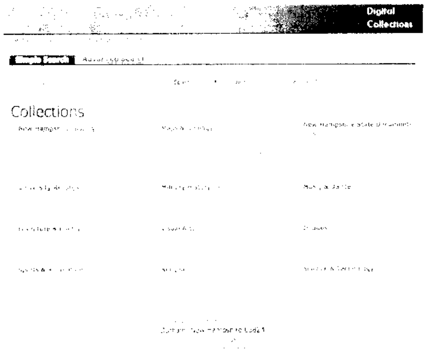
Refine search by Related Creator: Durham (N.H.: Town)

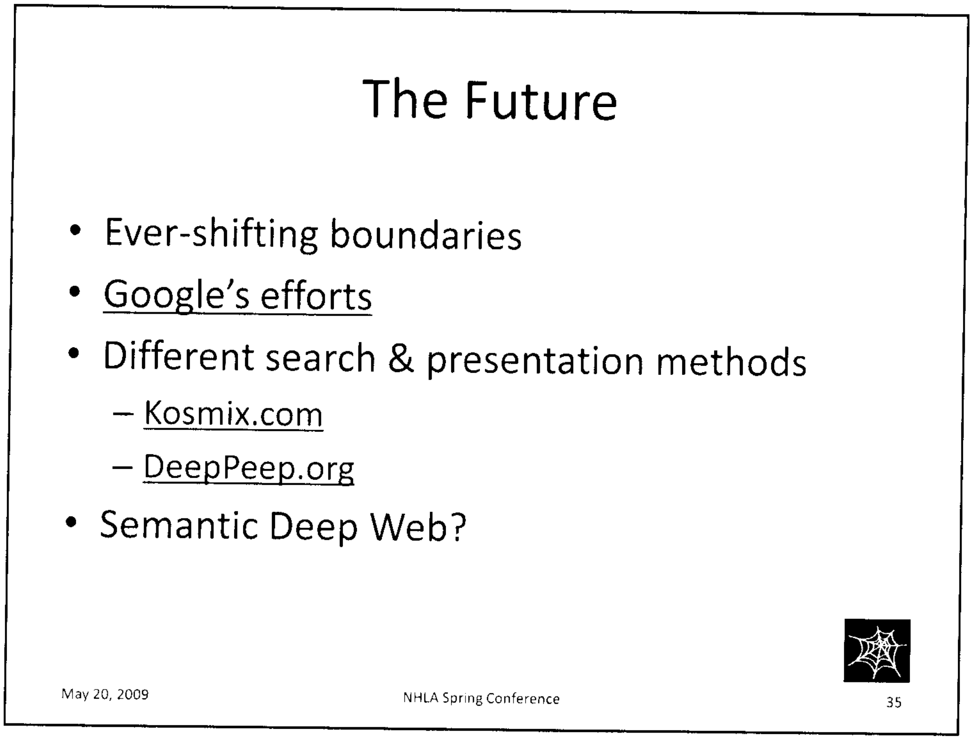
Sort by Date

Can view PDF, B&W PDF, full text, Flip Book

UNH Digital Collections Initiative Database

- http://www.library.unh.edu/diglib/
- Simple search searches metadata only
- Advanced search can search within full text
- Only format is PDF

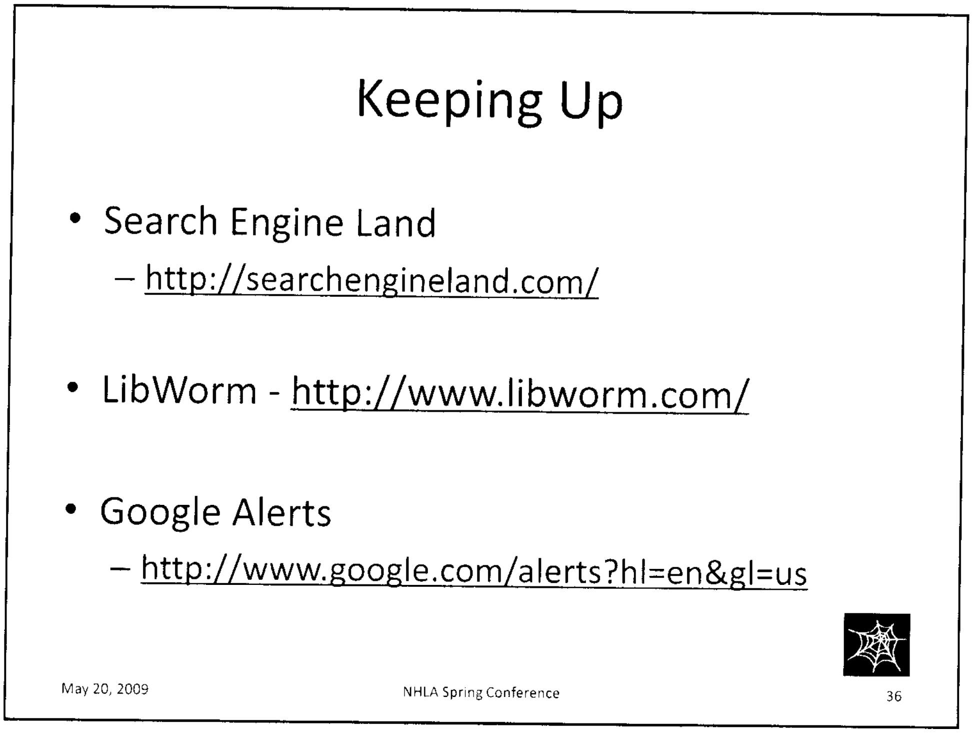




The Future

- Ever-shifting boundaries
- Google's efforts
- Different search & presentation methods
  - Kosmix.com
  - DeepPeep.org
- Semantic Deep Web?

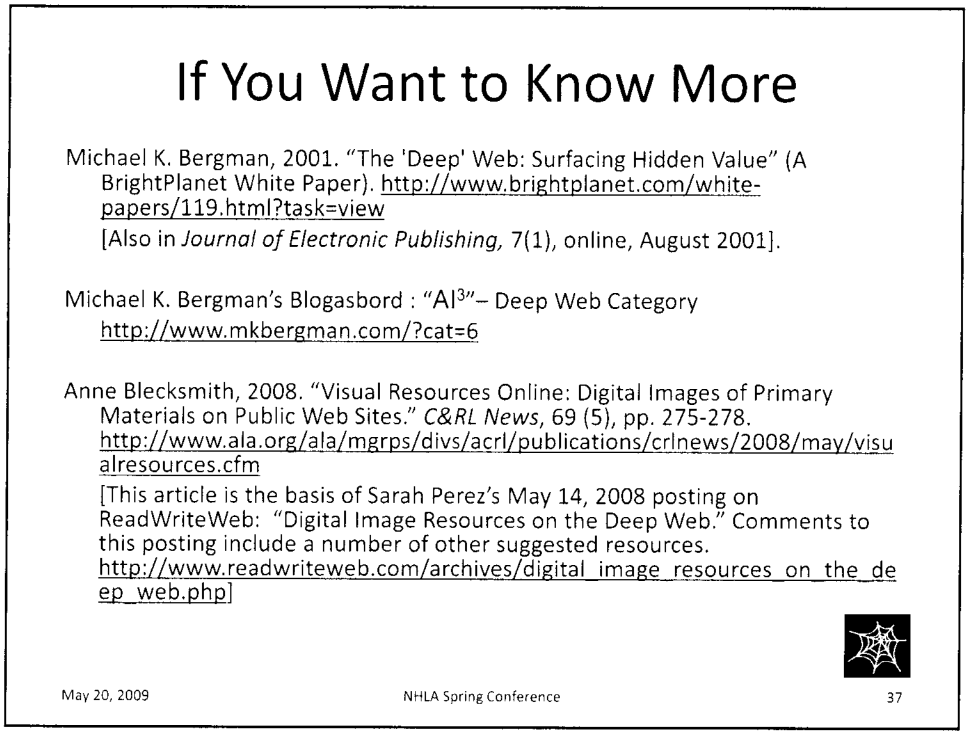
May 20, 2009 NHLA Spring Conference 35



Keeping Up

- Search Engine Land
  - http://searchengineland.com/
- LibWorm - http://www.libworm.com/
- Google Alerts
  - http://www.google.com/alerts?hl=en&gl=us

May 20, 2009 NHLA Spring Conference 36



If You Want to Know More

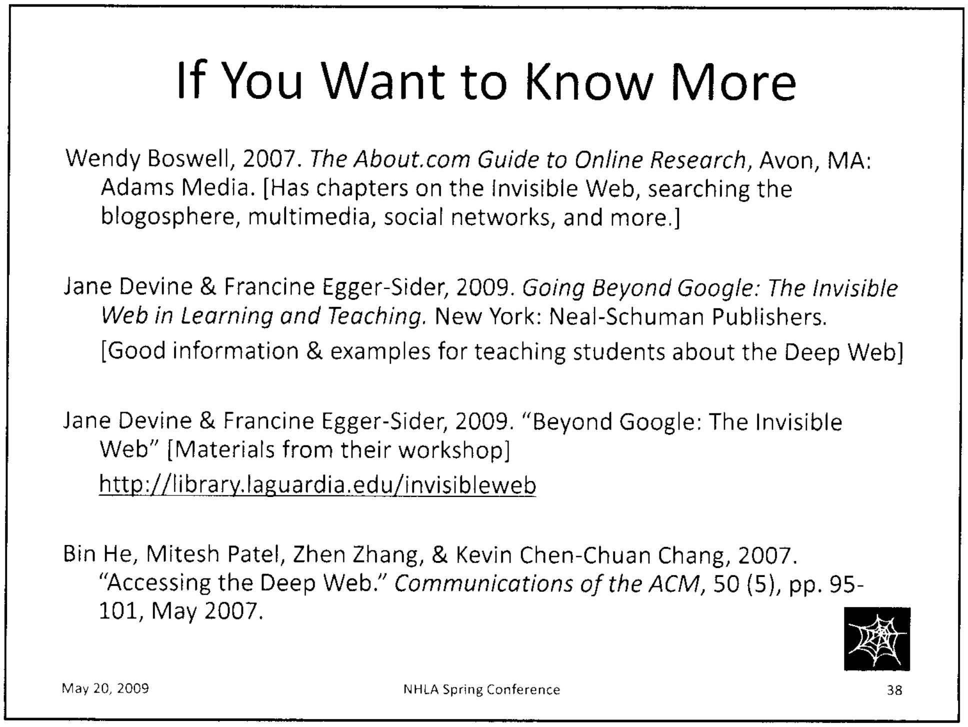
Michael K. Bergman, 2001. “The 'Deep' Web: Surfacing Hidden Value” (A BrightPlanet White Paper). http://www.brightplanet.com/white-papers/119.html?task=view

[Also in *Journal of Electronic Publishing,* 7(1), online, August 2001].

Michael K. Bergman’s Blogasbord : “AI^{3}”– Deep Web Category http://www.mkbergman.com/?cat=6

Anne Blecksmith, 2008. “Visual Resources Online: Digital Images of Primary Materials on Public Web Sites.” *C&RL News*, 69 (5), pp. 275-278. http://www.ala.org/ala/mgrps/divs/acrl/publications/crlnews/2008/may/visualresources.cfm

[This article is the basis of Sarah Perez’s May 14, 2008 posting on ReadWriteWeb: “Digital Image Resources on the Deep Web.” Comments to this posting include a number of other suggested resources. http://www.readwriteweb.com/archives/digital_image_resources_on_the_deep_web.php]



If You Want to Know More

Wendy Boswell, 2007. *The About.com Guide to Online Research*, Avon, MA: Adams Media. [Has chapters on the Invisible Web, searching the blogosphere, multimedia, social networks, and more.]

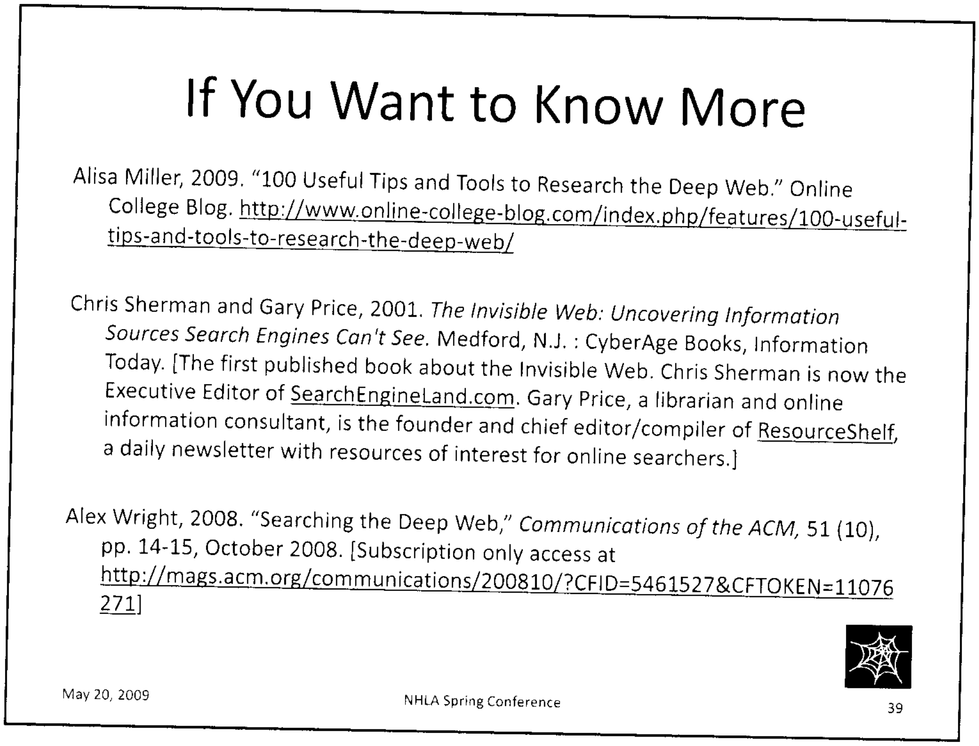
Jane Devine & Francine Egger-Sider, 2009. *Going Beyond Google: The Invisible Web in Learning and Teaching.* New York: Neal-Schuman Publishers.

[Good information & examples for teaching students about the Deep Web]

Jane Devine & Francine Egger-Sider, 2009. “Beyond Google: The Invisible Web” [Materials from their workshop]

http://library.laguardia.edu/invisibleweb

Bin He, Mitesh Patel, Zhen Zhang, & Kevin Chen-Chuan Chang, 2007. “Accessing the Deep Web.” *Communications of the ACM*, 50 (5), pp. 95-101, May 2007.



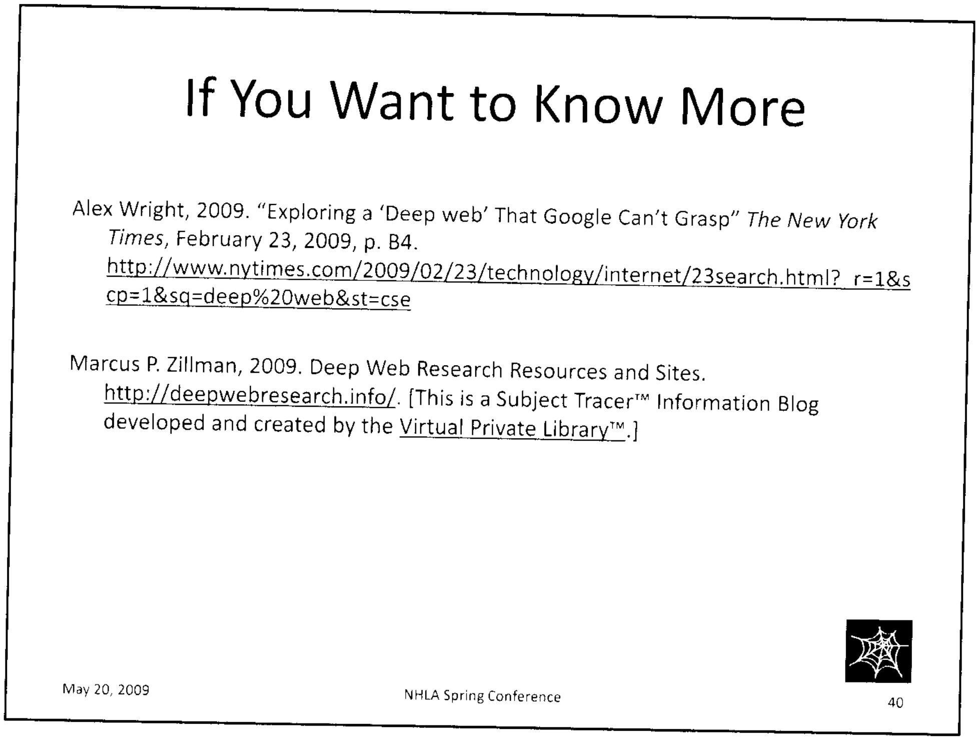
If You Want to Know More

Alisa Miller, 2009. "100 Useful Tips and Tools to Research the Deep Web." Online College Blog. http://www.online-college-blog.com/index.php/features/100-useful-tips-and-tools-to-research-the-deep-web/

Chris Sherman and Gary Price, 2001. *The Invisible Web: Uncovering Information Sources Search Engines Can't See.* Medford, N.J. : CyberAge Books, Information Today. [The first published book about the Invisible Web. Chris Sherman is now the Executive Editor of SearchEngineLand.com. Gary Price, a librarian and online information consultant, is the founder and chief editor/compiler of ResourceShelf, a daily newsletter with resources of interest for online searchers.]

Alex Wright, 2008. "Searching the Deep Web," *Communications of the ACM,* 51 (10), pp. 14-15, October 2008. [Subscription only access at http://mags.acm.org/communications/200810/?CFID=5461527&CFTOKEN=11076271]

May 20, 2009 NHLA Spring Conference 39



If You Want to Know More

Alex Wright, 2009. "Exploring a 'Deep web' That Google Can't Grasp" *The New York Times*, February 23, 2009, p. B4. http://www.nytimes.com/2009/02/23/technology/internet/23search.html?_r=1&scp=1&sq=deep%20web&st=cse

Marcus P. Zillman, 2009. Deep Web Research Resources and Sites. http://deepwebresearch.info/. [This is a Subject Tracer™ Information Blog developed and created by the Virtual Private Library™.]

May 20, 2009 NHLA Spring Conference 40

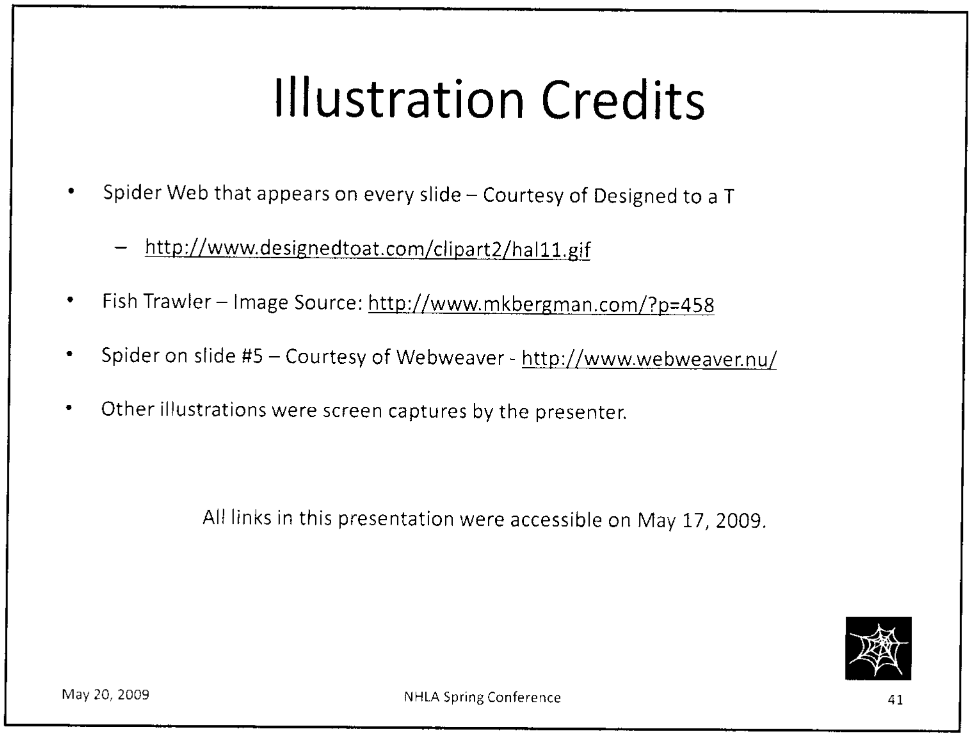


Illustration Credits

- Spider Web that appears on every slide – Courtesy of Designed to a T
  - http://www.designedtoat.com/clipart2/hal11.gif
- Fish Trawler – Image Source: http://www.mkbergman.com/?p=458
- Spider on slide #5 – Courtesy of Webweaver - http://www.webweaver.nu/
- Other illustrations were screen captures by the presenter.

All links in this presentation were accessible on May 17, 2009.