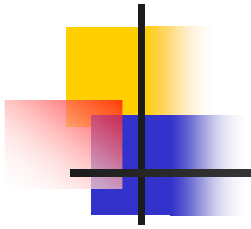


Live Web Search

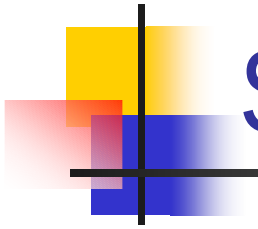
Mary Hodder
CEO: Bloqx, Inc.

mary@hodder.org
Blog: napsterization (napsterization.org/stories)

Live Web Buzz:

Doc Searls:

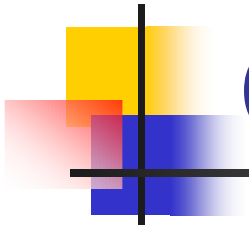
‘the live web’



Search: static web

- Google was created for the static web
  - Pagerank is a function of people's interest on one site, in another site, as expressed in links.
  - Google’s genius was equating those links to social gestures that had value.
  - Key word search returns the 'most relevant' results based upon whomever has the most pagerank.

Search for ‘mary hodder’ in Google



Yields:

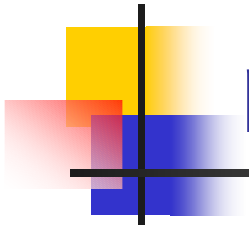
- ‘napsterization’
- ‘biplog’
- ‘speaking engagements’
- ‘bloqx’
- ‘hodder.org’

Search for ‘napsterization’ in Google

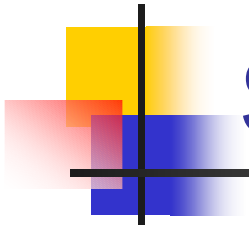
Yields:

- ‘mary hodder’
- The napsterization blog itself
- Uses of the word, napsterization

Search for ‘mary hodder’ or napsterization in blog search



- Yields key word matches
  - Reverse chron order
- Link search: yields all those who have linked to url
  - Reverse chron order
  - Can resort by Authority in some services



Static webpage search

- Static Webpages: Example webpage from a bank.
  - Incoming links 'define the bank' referring to it as an entity, but the bank doesn't change it’s pages.
  - Outgoing links from bank site rarely change.
  - Don’t really want sites like that to update a lot; useful in static nature

Why do we care about live web search?

- Marketing reasons: blog creators more likely to be:
  - Men: 57% are male
  - Young: 48% are under age 30
  - Broadband users: 70% have broadband at home
  - Internet veterans: 82% have been online for six years or more
  - Relatively well off financially: 42% live in households earning over $50,000
  - Well educated: 39% have college or graduate degrees
- They have characteristics of “opinion leaders”
- Their influence online is growing

- Source: Pew Internet & American Life Project, May 2, 2005

Live webpages: what and how

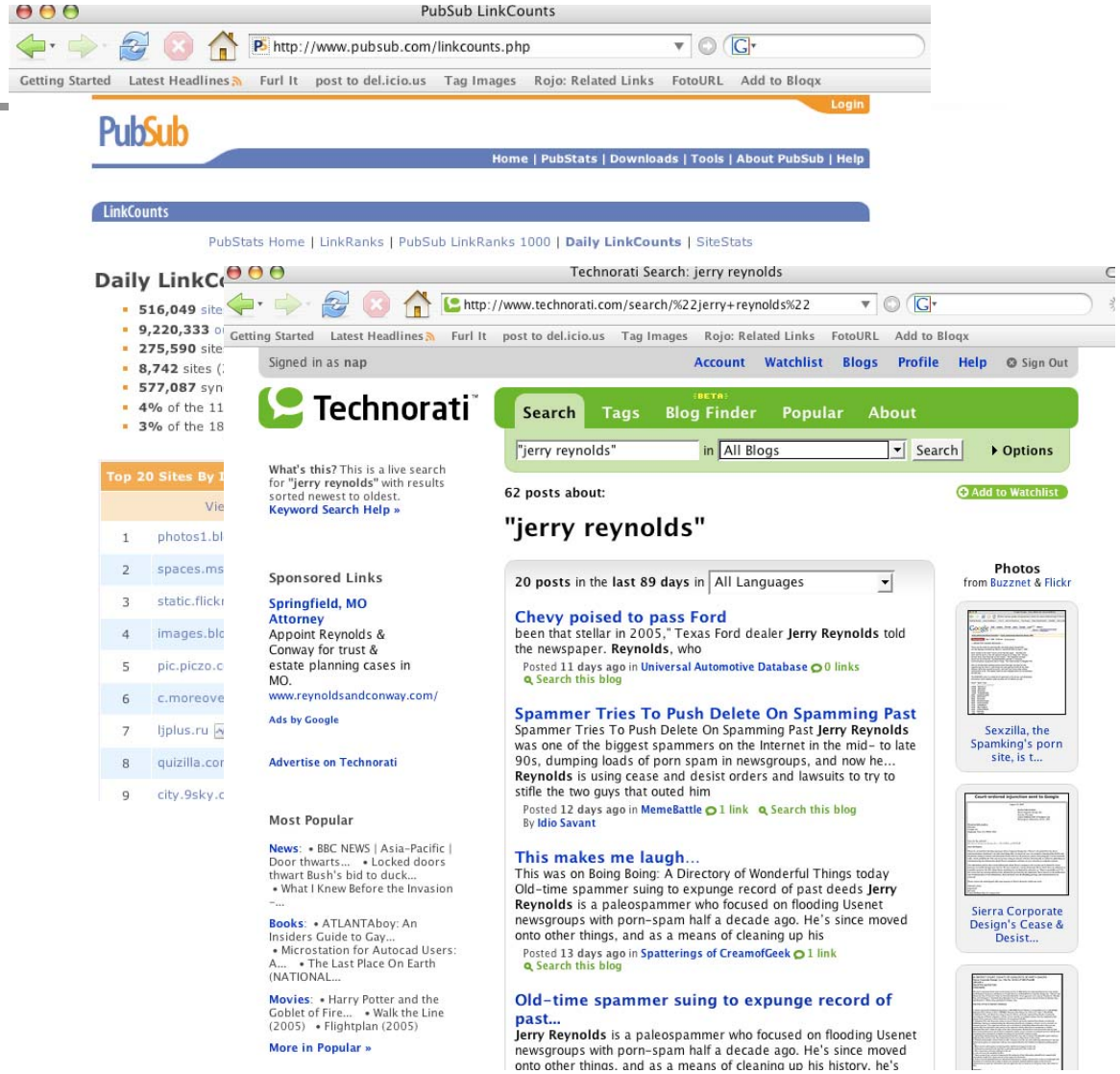
- Live web pages: rapidly updated pages such as Flickr photo sets, or blogs, often with lots of links and references in new posts.
- The constant activity produces many more outbound links and references, and many more inbound links, as conversations between creators appear in threads
  - Old blog entries scroll down and off blogs, but in Google sees them all as subpages
  - Page rank makes no distinction because it defines pages as 'any page' instead of 'top pages' as technorati, or pubsub do when they 'age' things out of the system..

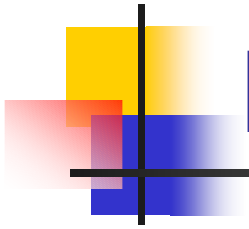
Metrics of Blog Search

Examples of those who rely on link counts:

- Technorati
- Pubsub
- Bloglines

“authority” vs. “influence”





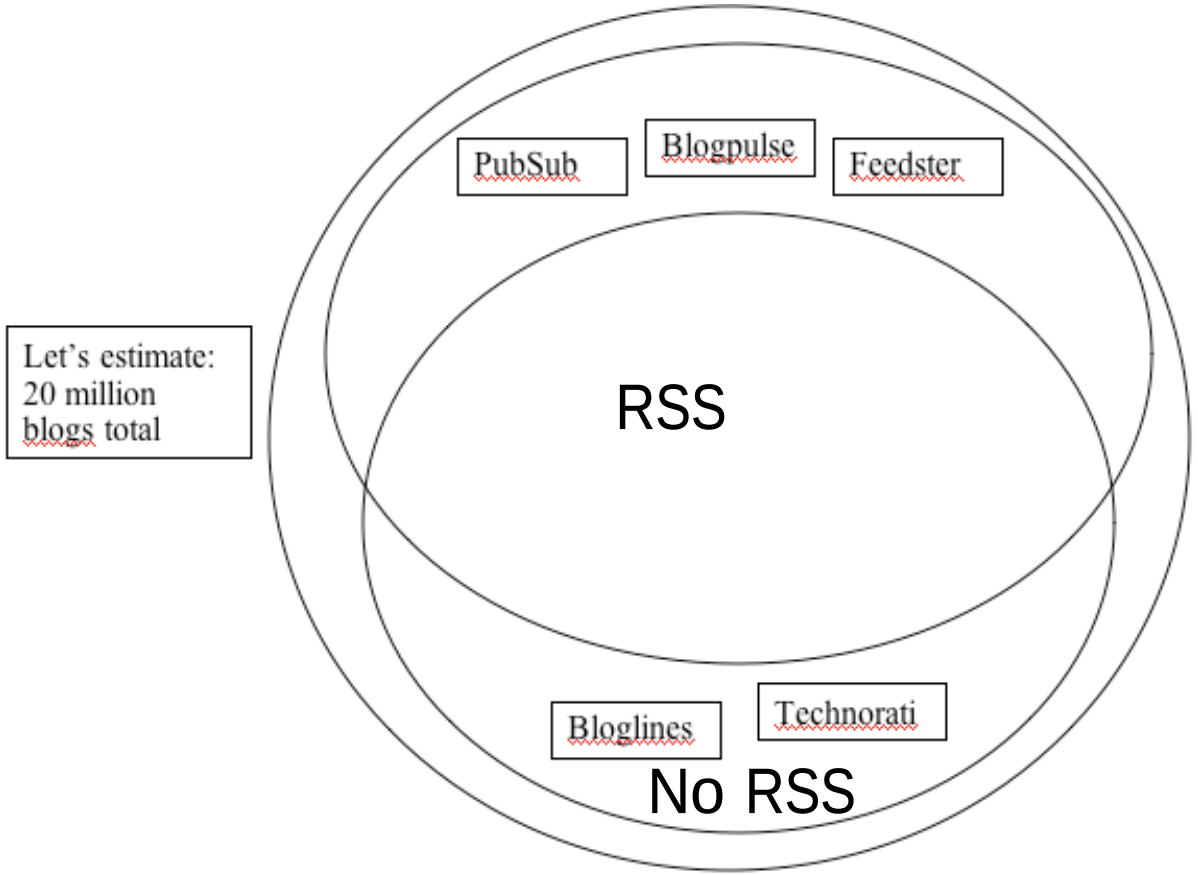
Live Web Examples

Bloglines: 880,456,867 Articles Indexed
Blogpulse: 18,447,338 feeds
Feedster: 18,047,721 feeds
Ice Rocket: no info
Pubsub: 18,549,449 sources, 11.8m active
Technorati: 21.5m blogs indexed

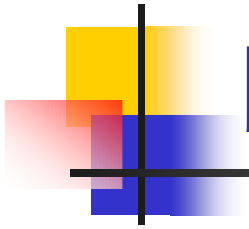
Del.icio.us: tags, urls and people feeds
Flickr: Photos, tags and sets/groups by 1.5m people

* As of 11/21/05

The Blogosphere: what are you searching?



Link metrics used to weigh bloggers



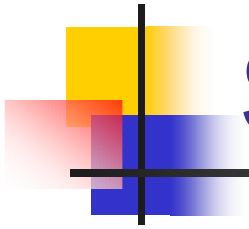
- Pubsub: Linkrank based on internal calculation of worth based on inbound links and those linkranks of those who link
- Technorati: count of all inbound link sources for last 6 months
- Bloglines: count of all links for all time

Subscription metrics

- Bloglines
  - Must subscribe to feed to see subscription numbers
- Feedburner
  - Must run your feed through Feedburner
  - Subscription numbers only available to feed owner
- Example: Boing Boing * as of August 2005
  - In Bloglines, 26,856 subscribers
  - In Feedburner, 1.2 million subscribers to
  - Difference: Feedburner shows all subscriptions where as Bloglines shows just those who use Bloglines reader and subscribe

Possible other metrics

- Inbound links: to post url
- Inbound links: to blog url
- Comments to posts
- Ratios within topic/post
- Blog server
- Subscribers to RSS feeds and blogs
- Incoming Traffic Links
- Outbound blogroll
- Tagged urls



Search results

- Returned in reverse chron order
- Technorati: can reorder by authority (defined as inbound link count)
- Sphere: can order by date or by relevance (metric based on links and key word analysis)

Results of mixing live web with Google:

- Google: has lack of awareness of time about publishing of site
  - Good: information continues to be available
  - Bad: liveness is lost
- Awareness of links and activity
- Result: Live web activity can overwhelm static web

Example: ‘Jerry Reynolds’

- Spammer from 1990’s
  - Largest Usenet spammer in 1997
  - Owner of two sites: netzilla and sexzilla
  - His spam helped, with other spam, to kill Usenet over several years
  - Spam fighters tried to stop it but failed
- Wants this information removed
  - Using C&Ds to remove this documentation from Google
- Why? Google owns Usenet postings.. They come up in searches

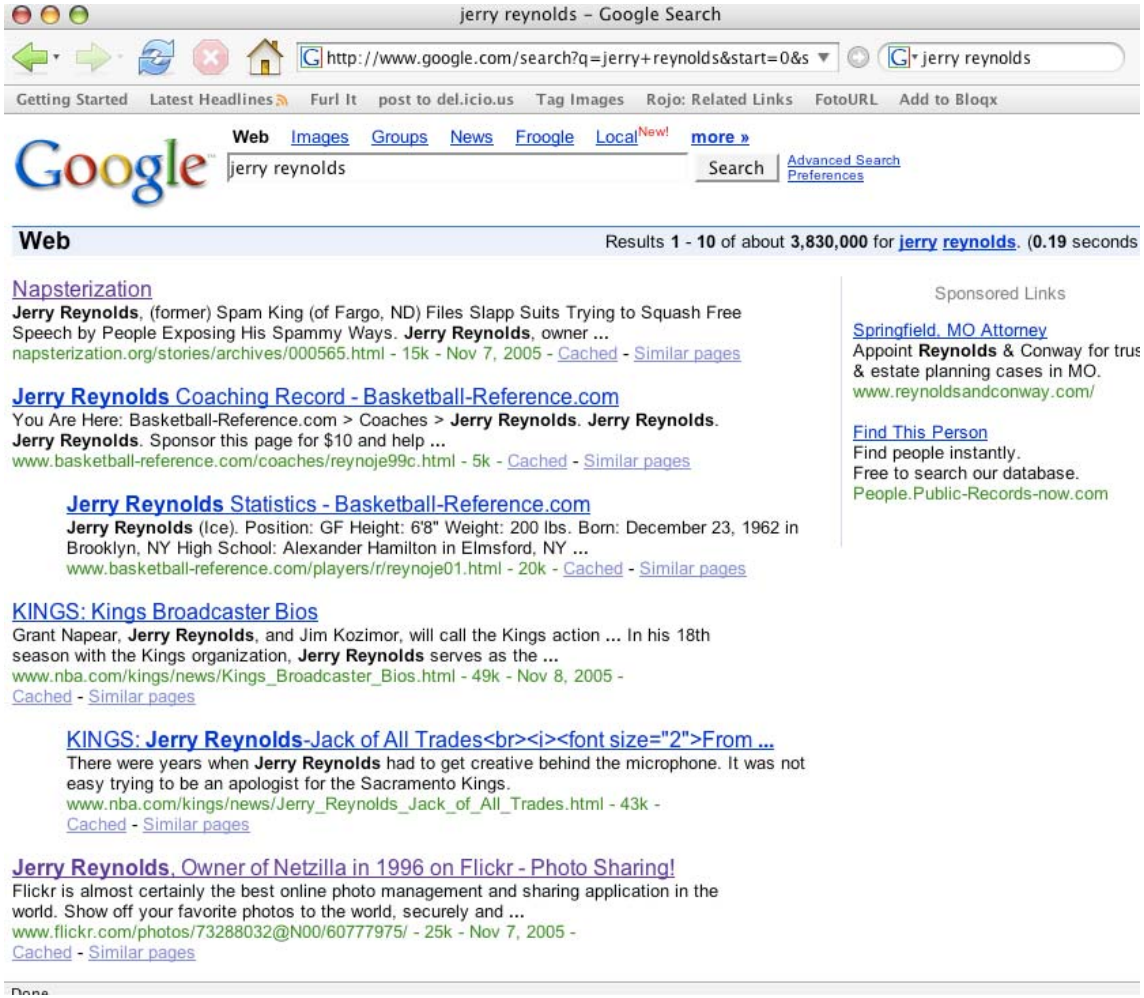
Before blogging about Jerry Reynolds

- 13 pages back in Google Results
- Unknown for years by anyone in technology journalism
  - Exception: one article in local paper in North Dakota

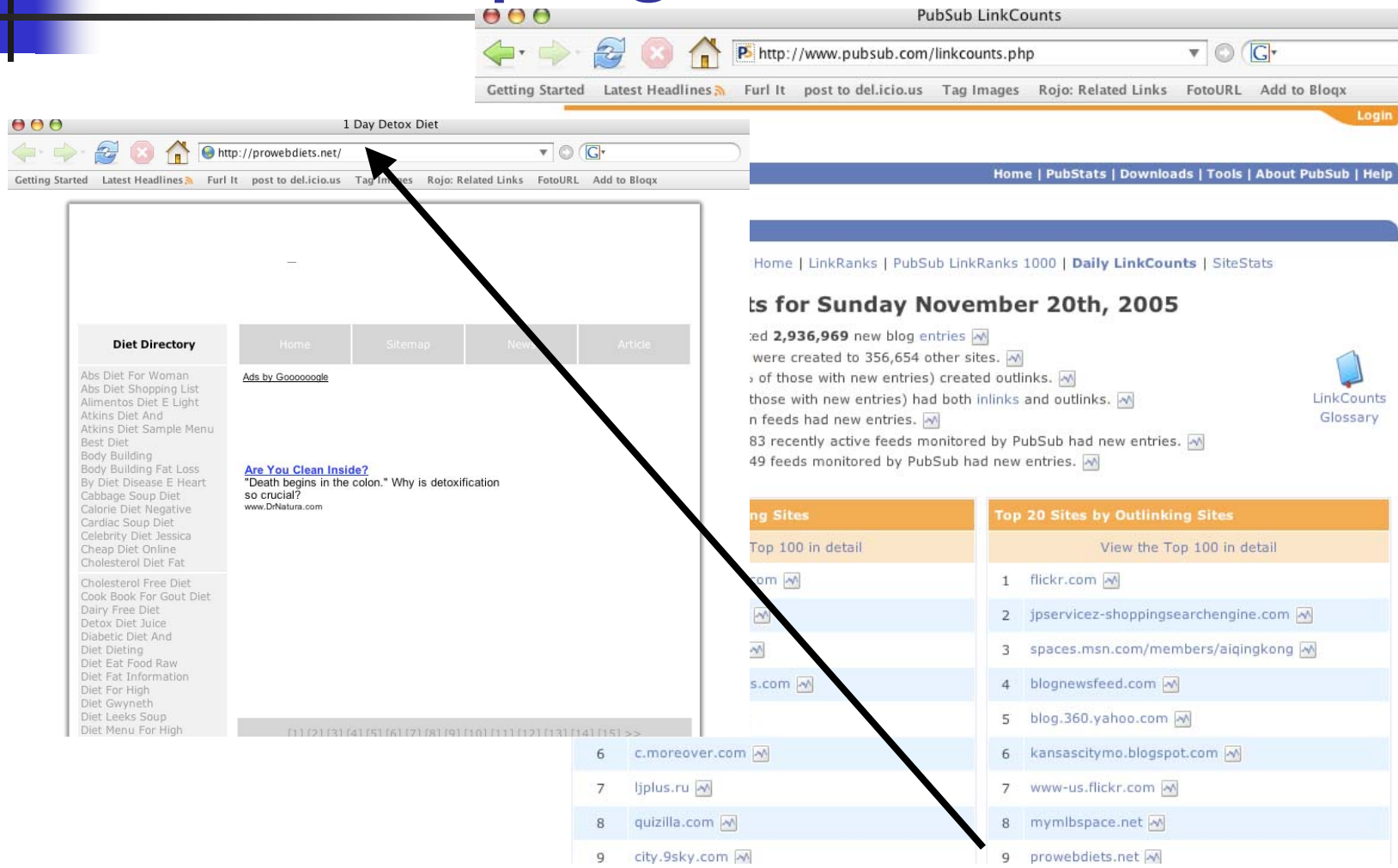
Blogging of Jerry Reynolds story

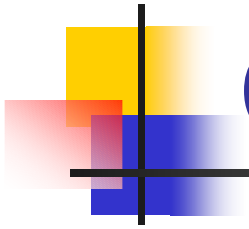
-BoingBoing
-Splatterings
-Memeblog

Post Blogging: #1 hit on Google



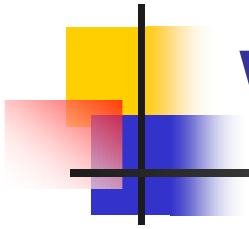
Abuse: Splogs





Google and Splogs

- Blogger
- Adsense
- Incentives to change?



What's needed

(Blog search is at the state of static web search in 1997)

- Sophisticated interfaces
- Topic browsing
- Sophisticated weighting tools
- Static web search adjustments

What's coming

- 'Data web' verses 'file web'
  - http://you.com/files/here.htm = data web
  - iTags: iname + tag + object = floating data
- Better search
  - Based on better live information
  - Based on data web?

Contact Info:

- Mary Hodder
- mary@hodder.org
- Blog: napsterization
- Http://napsterization.org/stories/