

Comparison of URL lookups for Bloglines, Blogpulse, Feedster, Pubsub and Technorati (by Mary Hodder, Napsterization)

Service	Information source¹ and method	Information they pull in: Post or Posts and Blogroll	Inbound link counts²	Inbound site or source counts³	Duration those links stay in count	Has links to one's own blog included	Duplicate posts show in link lookup results for Napsterization⁴	# of links and sources for Napsterization	Provide watchlist or alert to new inbound links for URL	Provides Link Rank for Blog based upon inbound links	Information Philosophy
Bloglines	RSS feeds and spider for html	Posts and blogroll	Yes	No	Forever	Yes	Yes (there were three of one post, two each of another)	1012 links	No	No	Has information going back to the beginning of their service. Also, only has blogs that are subscribed to by at least one person in a Bloglines newsreader.
Blogpulse	RSS feeds and spider for html	Just posts (they throw out blogroll data)	Yes	No	6 months	No	Yes, 2 of one post on the second page.	477 links on URL search, but profile notes 44 citations ⁵	Yes	Yes	Keeps information for 6 months.
Feedster	RSS feeds	Just posts or portion of post going through RSS feed	Yes	No	Unknown	No	Yes (there were two of one post on page one)	7 for nap/stories 351 for napsterization.org	Yes	No	Feedster had URL information going back to March of this year for napsterization.
Pubsub	RSS feeds	Just posts or portion of post going through RSS feed	Yes	Yes	30 days	No	No – though it's hard to prove because they just show link counts	76 links from 61 sources	Yes - Pubsub is really set up to provide subscriptions for searches	No	Pubsub does not keep historical information for search, but it does keep link rank data.
Technorati	RSS feeds and spider for html	Posts and blogroll, but ONLY data on front of blog is in link counts ⁶	Yes	Yes	While on front page of a blog	No	Yes (the first page of results showed 7 dups of one post; 2 of another for nap.org/stories)	271 links from 205 sites	Yes	Yes	Technorati sees the aging of information as a function of the front of a blog: if information is not on the front page, the information is 'old' and therefore falls out of Technorati link counts and url lookups.

¹ Companies that pull in RSS feeds only get the portion of the post that goes through the feed. However, some publishing companies, like Live Journal, provide a direct source, so for those published posts, the complete post is collected anyway. Companies that spider may get the complete post, but if they only spider the front page, and part of a post is in the “extended section” then the extended part may not be aggregated.

² Inbound link counts show all links to a blog, no matter where they are from (blogrolls or posts).

³ Inbound site or source totals count each blog once, no matter how many links from that blog, point to another blog. This includes blogrolls for Technorati. If a blogger links 10 times in any combination of blogroll and posts, the source count is 1. Pubsub aggregates those links and sources by day, so counting over the 30 days they show on their site is manual, but see this for a comparison: http://www.pubsub.com/site_stats.php?site=sifry.com%2Falerts

⁴ This happens for multiple reasons: post detection problems with spidering (for those companies that spider), companies pull in all the RSS feeds for a blog and therefore may get three or more posts that are the same, but come in different RSS feeds, and because bloggers resave posts as they update, it looks to aggregators of data as though a new post has appeared, or if a post has changed slightly, that the post is an new, and therefore, they want to collect and display all versions. Also, for this category, I tested my own blog at napsterization.org/stories/ because I already watch the inbound links and know the recent history, so it was an easy comparison and I could pick out duplicates easily.

⁵ Note that my initial search for the url string: napsterization.org/stories/ produced 9 results in Blogpulse. But they alerted me to the UI bug they have, where if the URL string includes <http://> in front of the URL, the results are more accurate.. including all the links they have. This is a bug they need to fix, but 477 results sounds much more like what I see in my Blogpulse search feed for napsterization.org/stories/.

⁶ For Technorati, if a blog post does not go through an RSS feed, and part of the post is accessible only as a background page through a ‘more’ or ‘extended post’ link, then Technorati does not spider the rest of the post. Technorati only spidered and aggregates the top page of a blog. Link counts only include this top home page data from blogs, though other searches do show data that has scrolled off the home page.