

Comparison of Keyword Searches 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	Key word search of historical info	Duration those keywords stay in results	Duplicate posts show in link lookup results for keywords²	# of keyword matches for Napsterization	Provide RSS watchlist or alert to new keywords	Information Philosophy
Bloglines	RSS feeds and spider for html	Posts data in full	Yes	Forever – but results show some lesser amt	Yes – more than half the posts were from my Napsterization blog; there were many more dups by my Nap blog and others)	85 Matches – ordered by date, most recent match: 12/12/04	No	Information goes back to the beginning of their service, but keywords searches only cover a portion of their database. Also, results only include blog posts from 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	6 months	Yes (about 10 dups in the first 100 posts)	205 matches	Yes	Keeps information for 6 months.
Feedster	RSS feeds	Just posts or portion of post going through RSS feed is collected; del.icio.us data and top down news also included	Yes	6 months	No - None in first 10 pages (about 100 results) for Napsterization	2,767 matches according to p1 of search results, but jumped around from 314 depending on the page ³	Yes	Have information in database since they started tracking, but they show about 6 months in results now. However, the design metric is to have all data available soon.. as they build out storage. Also mixes blog post, top down media and del.icio.us results together for searches.
Pubsub⁴	RSS feeds	Just posts or portion of post going through RSS feed	No	N/A	N/A	N/A	Yes - Pubsub is really set up to provide subscriptions for searches	Pubsub does not keep historical information for search, but is instead designed for keyword search feeds, which then pick up matches and send them through an RSS feed to you.
Technorati	RSS feeds and spider for html	Posts and blogroll keywords but ONLY data on front of blog ends up in search results ⁵	Yes	Since 10/04	Yes (a total of 12 dups in the first 100 posts and they had blogroll mentions mixed with post results)	323 matches	Yes	Technorati has key word results going back to October, 2004. It's an arbitrary date, but they continue to build the database from that date forward.

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 the direct source to search companies, so for those published posts, the complete post is collected anyway. Companies that spider may get the complete post, but not if they only spider the front page. Those parts of posts in the “extended section” may not be aggregated.

² Duplication of posts happens for multiple reasons: post detection problems with spidering (for those companies that spider), or 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 at least different, and therefore, they want to collect and display all versions. Also, for this category, I tested the word ‘napsterization’ because I already watch that and know the recent history, so it was an easy comparison and I could pick out duplicates easily.

³ Feedster explained that it uses the standard ‘estimation’ of the number of results, and while most search engines, Google included, maintain the same estimate across several pages, Feedster recalculates the estimates as each next page of search results are rendered, and that’s why the total changes (sometimes dramatically) from p1 to p2, and from p2 to p3, etc.

⁴ Pubsub is a feed subscription service, meaning that it is set up to create search feeds, where any matches going forward after the creation of the search are matched and sent to the user. They do not keep historical information, so not much of the historical keyword search comparison applies to their service.

⁵ 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 spiders and aggregates the top page of a blog. While link counts only include this top home page data from blogs, all other searches do show data that has scrolled off the blogger’s home page.