

Hot Weblog Crawling Action

An attempt to find as many active weblogs as possible, across all languages.

Latest Numbers

Here's a peek at our crawl status. The first number shows how many sites are now stored in our database. The second number is our best guess as to how many of these are weblogs. "Errors" and "Forbidden" refer to pages that gave us some kind of error, or were blocked by a robots.txt file. The blog queue is the list of sites we get from weblogs.com, the unknown queue is a list of top-level domains that we've found links to, but haven't visited yet.

Crawl Statistics

Status	Count	Definition
Queue	1631	Sites we know nothing about yet
Likely weblogs	700837	Visited sites we think are weblogs
Anglo weblogs	380636	Blogs that seem to be in English
Weblog Queue	1624	To-do list of known weblog sites

Report generated at 8:12 EST on July 31, 2003

285 sites crawled in the last five minutes (2906 last hour)

20 new blogs added in the last five minutes (355 last hour)
































205 known blogs updated in the last five minutes

Currently running 38 crawlers

Anglo Blogs

This is a provisional, running tally of authoring tools for weblogs written in English. Look further below for stats across all languages.






















Anglophone Weblogs: Authoring Tools

Type	Count	
Blogspot	135899	
LiveJournal	79489	
Diaryland	35428	
Movable Type	17420	
Pitas.com	16801	
Blogger	15151	
xanga	13212	
Radio UserLand	8904	
<i>Unknown (by URL)</i>	8553	
<i>Unknown (by RSS feed)</i>	5843	
Bloggningnetwork	2872	
Journalspace	2743	
<i>Unknown (by content)</i>	2361	
Tripod.com	2024	
b2	1983	
Manila	1673	
PHP-Nuke	1390	
Postnuke	1232	
Salon Radio Weblogs	1097	
Persianblog [IR]	1056	
Microsoft Visual Studio	1008	
Blog.PL [PL]	938	
G-Blog	727	
Blogger Pro	670	
pMachine	659	
Greymatter	527	
Blog-City	519	
Blog Studio	434	
Blogger [BR]	407	
Antville	362	
Nucleus	223	
skyblog	198	
Bloxom	142	
Weblogger [BR]	137	
e-blog.pl (polish)	116	
Blogsky [IR]	103	
CrimsonBlog	100	

Pivot	96	
Splinder [IT]	78	
Jevon.org	76	
diaryhub	59	
Blogging tool hompages	34	
Joueb [FR]	30	
Twin Cities Babelogue	25	
PsychoBlogger	18	
WebCrimson	17	

Why Don't They Speak English?

Here are comprehensive stats on authoring tools for languages other than English. Notice that not all languages are crawled at the same rate, so there are likely to be strange imbalances until we get into the millions or so. If you have questions about the crawl, or (better still) a list of known weblog URLs you'd like to contribute, please [contact me](#)

Non-English Weblogs: Authoring Tools		
Type	Count	
Blogspot	78654	
Blog.PL [PL]	51826	
Persianblog [IR]	31131	
Blogger [BR]	25427	
Pitas.com	13651	
Weblogger [BR]	12490	
Radio UserLand	11637	
<i>Unknown (by URL)</i>	8968	
LiveJournal	8356	
skyblog	7503	
Diaryland	7444	
Blig [BR]	5187	
Blogger	5069	
Movable Type	4690	
Splinder [IT]	3255	
<i>Unknown (by RSS feed)</i>	2598	
diaryhub	2320	
Manila	2298	
MonBlogue [FR]	2186	
PHP-Nuke	1780	
Microsoft Visual Studio	1475	

Tripod.com	1314	
<i>Unknown (by content)</i>	1247	
Blog Studio	1073	
Antville	1028	
Blogsky [IR]	954	
Joueb [FR]	856	
b2	777	
Postnuke	672	
Barrapunto [ES]	632	
U-Blog [FR]	431	
Salon Radio Weblogs	388	
Twoday [DE]	364	
e-blog.pl (polish)	355	
xanga	316	
Persianlog [IR]	312	
Pivot	222	
Blogalia	217	
pMachine	203	
Journalspace	180	
Nucleus	141	
Greymatter	116	
weblog.pl (polish)	112	
Blogger Pro	68	
Blog-City	66	
Bloggng tool hompages	29	
CrimsonBlog	25	
Bloggngnetwork	18	
Bloxom	12	
Popular template site	12	
Jevon.org	11	
G-Blog	10	
WebCrimson	5	
slogger	4	
Twin Cities Babelogue	2	
PsychoBlogger	2	
Textpattern	1	
land down under	1	

Language Distribution

Here is a language breakdown of the blogs crawled so far. Language identification is done using TextCat, a little program that looks at the statistical distribution of three-letter clusters. Language data may be skewed as we stagger around the world looking for blogs, but should even out over time.

Language Distribution		
Language	Count	
English	380637	
Too_short	120241	
Portuguese	55858	
Polish	42714	
Farsi	29004	
French	11790	
Spanish	10037	
German	8158	
Italian	7165	
Dutch	3853	
Icelandic	3634	
??	3138	
Chinese-big5	3137	
Catalan	3131	
Thai	2373	
Indonesian	1909	
Chinese-gb2312	1372	
Malay	1324	
Russian	922	
Latin	676	