The ultimate guide to bot herding and spider wrangling

This is Part 1 of a three-part series.

We generally think about search engine optimization in relation to humans: What queries are my customers using?

How can I get more bloggers to link to me?

How can I get people to stay longer on my site?

How can I add more value to my customers' lives and

businesses?

This is how it should be.

But even though we live in a world that is increasingly affected by non-human actors like machines, artificial intelligence (AI) and algorithms, we often forget a large part of optimizing a website has nothing to do with people at all.

In fact, many of the website visitors we need to please are actually robots, and we ignore them at our peril!

What is a bot, anyway?

A bot (also known as a spider or crawler) is simply a piece of software that Google (or another company) uses to scour the web and gather information or perform automated tasks.

The term “bot” or “spider” is slightly misleading, as it suggests some level of intelligence. In reality, these crawlers aren’t really doing much analysis. The bots aren’t ascertaining the quality of your content; that’s not their job. They simply follow links around the web while siphoning up content and code, which they deliver to other algorithms for indexing.

These algorithms then take the information the crawler has gathered and store it in a massive, distributed database called the index. When you type a keyword into a search engine, it is this database you are searching.

Other algorithms apply various rules to evaluate the content in the database and decide where a universal resource locator (URL) should be placed in the rankings for a particular search term. The analysis includes such things as where the highly related keywords appear on a page, the quantity and quality of the backlinks and the overall content quality.

By now, you’re probably getting the gist of why optimizing for

bots is important.

While the crawler doesn't decide whether your site will appear in search results, if it can't gather all the information it needs, then your chances of ranking are pretty slim!

So, how do you wrangle all those crawlers and guide them to where they need to be? And how do you give them exactly what they're looking for?

First things first: Understanding crawl budget

If you want to optimize your site for bots, you first need to understand how they operate. That's where your "crawl budget" comes in.

Crawl budget is a term search engine optimization specialists (SEOs) developed to describe the resources a search engine allocates to crawl a given site. Essentially, the more important a search engine deems your site, the more resources it will assign to crawling it, and the higher your crawl budget.

While many commentators have tried to come up with a precise way to calculate crawl budget, there is really no way to put a concrete number on it.

After the term became popular, Google weighed in with an explanation of what crawl budget means for Googlebot. They emphasize two major factors that make up your crawl budget:

- **Crawl rate limit:** The rate at which Googlebot can crawl a site without degrading your users' experience (as determined by your server capacity and so on).
- **Crawl demand:** Based on the popularity of a particular URL, as well as how "stale" the content at that URL is in Google's index. The more popular a URL, the higher

the demand, and the more it’s updated, the more often Google needs to crawl it.

In other words, your crawl budget will be affected by a number of factors, including how much traffic you get, the ease with which a search engine can crawl your site, your page speed, page size (bandwidth use), how often you update your site, the ratio of meaningful to meaningless URLs and so on.

To get an idea of how often Googlebot crawls your site, simply head over to the “Crawl: Crawl Stats” section of Google Search Console. These charts/graphs are provided for free from Google and indeed, they are helpful, but they provide a woefully incomplete picture of bot activity on your site.

Ideally, you should analyze your server log files with a program like OnCrawl or Screaming Frog Log Analyser.

It’s important to bear in mind that Google Search Console (GSC) is *not* a server log analyzer. In other words, there is no capability for webmasters to upload server logs to GSC for analysis of *all* bot visits, including Bingbot.

There are a few major things to consider when optimizing your crawl budget:

- **The frequency of site updates.** If you run a blog that’s updated once a month, don’t expect Google to place a high priority on crawling your site. On the other hand, high-profile URLs with a high frequency of updates (like HuffPost’s home page, for example) might be crawled every few minutes. If you want Googlebot to crawl your

site more often, feed it content more frequently.

- **Host load.** While Google wants to crawl your site regularly, it also doesn't want to disrupt your users' browsing experience. A high frequency of crawls can place a heavy load on your servers. Generally, sites with limited capacity (such as those on shared hosting) or unusually large page weights are crawled less often.
- **Page speed.** Slow load time can affect your rankings and drive away users. It also deters crawlers that need to gather information quickly. Slow page load times can cause bots to hit their crawl rate limit quickly and move on to other sites.
- **Crawl errors**. Problems like server timeouts, 500 server errors and other server availability issues can slow bots down or even prevent them from crawling your site altogether. In order to check for errors, you should use a combination of tools, such as Google Search Console, Deep Crawl or Screaming Frog SEO Spider (not to be confused with Screaming Frog Log Analyser). Cross-reference reports, and don't rely on one tool exclusively, as you may miss important errors.

This ends Part 1 of our three-part series: The Ultimate Guide to Bot Herding and Spider Wrangling. In Part 2, we'll learn how to let search engines know what's important on our webpages and look at common coding issues. Stay tuned.

Opinions expressed in this article are those of the guest author and not necessarily Search Engine Land. Staff authors are listed here.

About The Author

Stephan Spencer is the creator of the 3-day immersive SEO seminar Traffic Control; an author of the O'Reilly books *The Art of SEO*, *Google Power Search*, and *Social eCommerce*; founder of the SEO agency Netconcepts (acquired in 2010); inventor of the SEO proxy technology GravityStream; and the host of two podcast shows The Optimized Geek and Marketing Speak.
http://feeds.searchengineland.com/~r/searchengineland/~3/Lv3mUZf_gZM/ultimate-guide-bot-herding-spider-wrangling-293284