morevisibility LEADERS IN SEARCH ENGINE MARKETING & OPTIMIZATION

Second Generation Search Engine Submissions – A Guide to What You Need to Know

Written by: Joe Laratro

Introduction

Which came first the Search Engine or the Directory Submission? Technically, that is a trick question because Directories most likely came first. The early Search Engines crawled the Web to find links. Crawling was inefficient because of the limitations of the crawlers (aka spiders). These Search Engines also depended on manual submissions by Webmasters to find new Web sites. Search Engine submissions have evolved tremendously over the last decade. What started out as a simple “Submit URL” that accepted domains has become a registration process where the Search Engines interact with Webmasters providing valuable information and insight into registered Web sites. This process is now called **Second Generation Submissions**.

First Generation Submissions

Search Engine submissions have a colorful past. There have been many programs to assist Webmasters in the submission process. While these programs could be used in a positive way (one submission per domain per engine), they could also be abused to send mass submissions to Search Engines. Mass Submissions, the automated process of sending every possible Web site URL in constant intervals, wasted the bandwidth of Engines and devalued potentially important submissions.

Engineers that worked for the Search Engines were constantly finding news ways to prevent automated submissions and keep the process ethical. Discovering new and valuable content was still one of the top goals, but Spam fighting became a full time job. Search Engines created Spam editor positions to help keep their results as relevant as possible. Spam editors teamed up with the Search Engineers to block automated submissions. They could identify IP addresses that were over submitting and flag sites that were being over submitted. *At one point there was even a penalty for using the free submit option*. It was more desirable for a Web site to be found through a Web crawl than because of a submission. One of the most successful innovations was the image verification code to block auto submissions.

Automated programs did not have the ability to read graphic files. Therefore codes were entered into graphic files and displayed to ensure human interaction with the submission pages. These images were advanced enough to include noise around the text to prevent Optical Character Recognition (OCR) software. For the most part automated submissions were dealt a swift and severe blow.

Directory Submissions

Directories were on many Webmasters checklists for submissions. This was certainly an important step. In fact a good Open Directory Project listing (Dmoz.org) was (and still is) a key to an instant Web presence. The Yahoo Directory used to be the first line of results in that engine, while Looksmart's Directory was the first line of results on MSN. There are many other directories that were (and still are) important for listings and traffic generation.

Directory submissions were less likely to be victims of submission spam, because they were human edited. They most closely resemble a standard Yellow page type of listing. Web sites typically had a single, one-category listing for the main domain. Some automated submission programs could submit to directories, but would not do so frequently or submit more than the home page. Directories started to monetize the submissions which helped pay for the editors to review them and weed out spammy sites that would not pay the submission fees.

These submissions were also different in other ways. A directory submission was more like a Web site registration. They collected much more information about the Web site: suggested title, suggested description, suggested keywords, category, location, email, and other fields.

XML Feeds

Another solution for indexing, submission, and spam problems was the innovation of XML Feeds. In March 2001, Inktomi first introduced a feed program for large dynamic Web sites that had difficulty getting the majority of their pages indexed. Inktomi's XML feed created a submission channel that plugged the database end of a Web site directly into the Search Engine's index. The program went a step further and setup a mechanism for smaller Web sites to "pay for inclusion" into Inktomi. Web companies could pay a per-URL fee to guarantee that they were crawled and included in the index. The larger program was based on CPC fees. The XML feeds followed the same path for monetization that Directories carved.

Paid Search Engine feeds seemed like the answer to many a Webmaster and Search Marketer's woes. Varying programs were implemented by all of the major Search Engines except Google. Through industry consolidations, the major XML programs have been reduced to just Yahoo's Search Submit and Search Submit Pro programs. They can still be a valuable source of traffic with guaranteed Search Engine Index inclusion.

Second Generation Submissions

There are several factors that led to the improved relationship and open communication between the major Search Engines and Webmasters. Webmaster World and Search Engine Watch were two very popular Web sites / forums for the online marketing community. GoogleGuy joined Webmaster World as a representative of Google to answer the public's questions. The advent of Search Engine Conferences brought Webmasters and Search Engines closer. Interestingly, the Search Engine algorithms were getting so advanced that many spammers were switching sides and abiding by the rules (Webmaster Guidelines).

When Google released the beta for Google Sitemaps in mid-2004, a new day was born for Search Engine Submissions. Google devised a free XML sitemap protocol for submitting a site to their index. This format was different than paid submission programs because there was no guarantee of indexing. Yet, it was free and the submission variables were significantly different. The XML sitemap was important, but it was not the cornerstone breakthrough of the Second Generation Submission technique. The Google Webmaster Tools interface, with site registration and insight into how the engine sees Web sites, was the true leap forward.

Google Webmaster Tools and Google Webmaster Tools Blogs

Through Google's Webmaster Tools, users are empowered with more information about the indexing of their Web site than ever before. There are also options to control certain aspects of the Web site that could not be previously communicated directly to the Search Engine.

This section of the paper will examine the four main categories of Google Webmaster Tools and their respective functionality and reports.

To get started, a Google Account is required. Then the Web site's domain name is added to the interface. The Web site has to be authenticated with Google in order to access the really good details. This process prevents competitors from gaining access to sensitive data. The site verification process can be completed by uploading a blank, but uniquely named file to the root of the domain, or a custom coded meta tag can be added to the index page.

Diagnostic – This is the first section a verified Webmaster would arrive at after signing into the domain. It includes valuable summary information and access to the Webmaster Tools.

> Summary – This report supplies the date of the last time Google accessed (crawled) the Web site. It is also the location for any spam penalty notifications. The penalty notification takes the guess work out of traffic losses from Google. If there is a penalty, a "reinclusion request" link will be available.
>
> Crawl Errors – This section will let a Webmaster know if there are any crawling issues found by Google's spiders.
>
> > Web Crawl – This report details the HTTP errors, pages not found, URLs not followed, URLs restricted by robots.txt, URLs that timed out, and unreachable URLs. This is great technical information about problems that can be very easily corrected.

Mobile Web – Google has special crawlers for mobile Web sites. They also have a separate site map submission for Mobile versions of Web sites. This report details any problems it finds in CTML, WML, and XHTML.

Tools – This section empowers Webmasters, Online Marketers, or Site Owners with tools that provide information directly to Google's Spiders.

Robots.txt Analysis – This report provides the contents of the robots.txt if it exists. It also has a tool that tests additions to the robots.txt file against sample pages. This is a great way to live-test changes to the robots.txt before actually implementing the changes. It can be used to test if certain areas of the Web site are blocked or if variable uses of URLs are being blocked (most commonly to prevent duplicate content).

Manage Site Verification – This is the area where Google supplies the customized verification code and method of uploading a file or adding a meta tag.

Crawl Rate – Google provides data about how frequently their spiders crawl the site and the amounts of data they collect. A Webmaster can choose to slow down the crawling speed if it is causing a performance drain on their Web servers. A crawl delay tag can also be added to the robots.txt file.

Preferred Domain – Most Web sites are setup with the "www." being a canonical name of their domain. This means that without a 301 redirect on the root of a domain, Search Engines would see http://domain.com and http://www.domain.com to be duplicates of each other. With preferred domain, a Webmaster can choose all results to default to the "www." or non "www." version of the domain.

Enhanced Image Search – This is an opt-in program offered by Google to enhance their image search engine with image content from registered Web sites.

URL Removals – This is a tool to allow Webmasters to quickly remove content that is displaying in Google's search results that is undesirable. This content could include out-dated pages, duplicate pages, pages with sensitive information, or a myriad of other types of pages that should not have been indexed.

Statistics – This section provides reporting on site information that was previously unknown to Webmasters. This information sheds light onto how Google sees the Web site.

Crawl Stats – This report focuses on part of Google's search algorithm known as PageRank. It gives quick information about how PageRank is divided across the domain and what the highest ranking page is.

Query Stats – This report displays the keywords or phrases that the Web site most frequently returns in search queries. It goes a step further by also displaying which keywords or phrases are actually driving Web traffic. Comparing words driving the traffic to words that are just driving impressions can be very useful for Search Engine Optimization efforts.

Page Analysis – These reports are also very useful for SEO. Data includes keywords and phrases used in external links to the site, keywords recognized in the content by density and other factors, and Web page types and encodings.

Index Stats – Google provides a menu of standard search operators like: "site:domain.com"; "link:domain.com"; "cache:domain.com"; "info:domain.com"; and "related:domain.com".

Links – Several years ago, Google changed the public results for the search operator “link:”. The command was being exploited by competitors and spammers to research successful Web sites and their back links. Through Google Webmaster Tools the engine has not only re-opened the full backlink information, but they now provide even more information and access to that information.

> External Links – This report is the holy grail of Google information for Webmasters. It includes how many links are counted toward each page, where the links are from, and what text is used in the links. As is true with most of the Webmaster Tools’ reports, this information can be downloaded and archived.
>
> Internal Links – This report shows the internal popularity of linking within the domain. It can be very helpful for site architecture and linking.

Sitemaps – Google Webmaster Tools provides 95% of the covered functionality without uploading a sitemap.xml file. The sitemap file is a list of all URLs existing within a Web site. Thanks to the sitemap protocol, there is an approved set of attributes that can be used in external sitemap files. These files are different then a typical HTML Sitemap (also referred to as internal sitemaps), as they are only used, by the engines, on the backend of a Web site and would not be viewed by regular Web users.

> Sitemap Uploads – Once a sitemap.xml file has been created it should be registered with Google through this interface. The file should be updated any time new content is added to the site, or there are changes to the site’s architecture. Web sites may have multiple sitemap.xml files based on content, and they may have separate files for mobile content. Sitemap files may be zipped to reduce server load. Google will report on the frequency of the sitemap being indexed, the number of URLs, and if there are any errors.

Sitemap.org Fields – For the most up-to-date sitemap information please visit http://www.sitemaps.org/protocol.php. These are the fields that may be used in a sitemap file:

Urlset – Start of the file
Url – Start of URL information
Loc – URL of the Web page
Lastmod – Date file was last changed
Changefreq – Frequency of the Web page updating (hourly, daily, monthly, etc)
Priority – Subjective user provided importance of the page (10 → 1)

Sitemap Generation Programs – Sitemaps can be manually compiled, exported from a database, or generated from a program. Sitemaps are only a suggestion for Search Engines. Their use does not guarantee that the Web pages will be crawled or included in search databases. Regardless, if external sitemaps are going to be used, they should be built with care. They should be an honest representation of the Web site and be useful for crawlers. Google provides a list of approved third party tools for sitemap generation at: http://code.google.com/sm_thirdparty.html. The Gsitecrawler is a very good option that includes other useful tools such as a duplicate content checker, meta tag export tool, and a very good dynamic crawler.

Yahoo Site Explorer and Site Explorer Blog

About the time Google was releasing its Sitemap Program, Yahoo started shifting searches with the “site:” operator to their Site Explorer system. This move shifted savvy searchers and Web site researchers off of their standard search platform. Yahoo’s Site Explorer Tool was fairly simple at first. It would allow anyone to track any site without verification. It also allowed for a channel to submit a “urllist.txt” file that could contain a list of every possible URL within the domain without any attributes.

Authentication – Yahoo allows users to authenticate their Web site by using an authentication key either as the name of a blank html file, or as a custom meta tag. Once a site is authorized, even more data is available. Yahoo accepts urllist.txt, xml sitemaps that fit the sitemap protocol, and sitemaps for mobile content.

Explore – The current version of the Yahoo Site Explorer is simpler than Google's Webmaster Tools. Yahoo has been continuously improving it and adding new features. Its Blog is a direct communication channel for the latest enhancements. The explore function gives wonderful insight into how Yahoo's spiders see the Web site.

Pages – The Site Explorer details indexed page information and can be sorted, restricted by sub-domain, and exported.

Inlinks – This report provides details about backlinks and can be sorted, filtered by domains, filtered by pages linked to, and exported. A nice exercise for SEO would be to compare the results of Google and Yahoo and compare and contrast the backlinks that the two engines recognize.

Delete URL/Path – Yahoo provides a tool that can swiftly remove up to 25 URLs from their database. The types of pages that may want to be removed are discussed in the Google Webmaster Tools section.

Future Adopters of Second Generation Submissions

MSN and Ask already announced that they have agreed to the sitemap protocol and will build programs to accept XML sitemaps. MSN still has a free "add URL" option, but Ask does not, instead it relies on Web crawls to find new information. It would be beneficial for both companies to follow Google and Yahoo's leads and open submission programs that could allow two way communications between Search Engines and Webmasters. The sitemap protocol is not limited to the Major Search Engines. It could be used by any Web site that is interested in crawling the Web for information.

Conclusion

If your Web site is not already registered with Google and Yahoo, now is the time to do it. They continually enhance the reports and tools available to their registered sites. The ability for the Search Engines to have direct contact to Webmasters or marketers is invaluable. If a Web site has been flagged for a Spam penalty within Google, the Webmaster Tools facilitate correction and resubmission. The reports are split between technical information and data that can be used on the marketing side. At a minimum, backlinks and pages indexed should be benchmarked for Web site growth. If a Web site is going to be redesigned, it is important to know what pages have inbound links. Then those pages can be saved, or have individual 301 redirects (a server-side message that says the content has been permanently moved) to the new corresponding page without losing link value. There are many more reasons to use these programs. Webmasters and marketers dreamed of having this data at their fingertips just a few years ago. Now the dream is a reality. The more these programs are used, and the more feedback that is provided to the Search Engines for improvements, the better the Web will be tomorrow for finding the Web sites that count.