# Making the Most of Your Content

*A Publisher's Guide to the Web*

Google™

## Contents

## Introduction

If you're looking for visibility, the Internet is the place to be. Just ask any advertiser who has increased sales using online ads, a blogger whose popularity has led to a book deal, or a newspaper that has expanded its readership to an international audience thanks to increased traffic.

At Google we're frequently asked how web search works, and how web publishers can maximise their visibility on the Internet.

We've written this short booklet to help you understand how a search engine 'sees' your content, and how you can best tailor your web presence to ensure that what you want to be found is found – and what you want to keep hidden, stays hidden.

From webmaster tips and online tools, to a step-by-step guide to frequently asked questions, this booklet is geared towards small web publishers as well as owners of large sites.

Just as the Internet itself has evolved dramatically in the past decade, so has Google's own approach to web search and its relationship with site owners. We've created numerous tools to help webmasters maximise the visibility of their content, as well as control how their web pages are indexed. But there's always more we can do. We hope that this booklet will encourage you to give us feedback and let us know what we can do to make the web an even better place for both searchers and publishers.

*- The Google Webmaster Team*

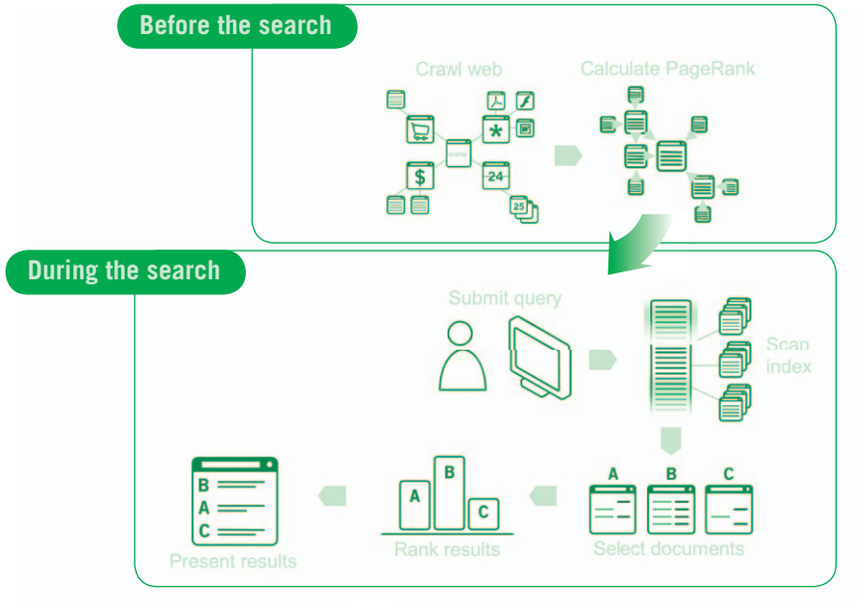
## A brief overview of web search: *How it works*

In the most simple of terms, you could think of the web as a very large book, with an impressive index telling you exactly where everything is located.

Google has a set of computers – the 'Googlebot' – that are continually 'crawling' (browsing) billions of pages on the web. This crawl process is algorithmic: computer programs determine which sites to crawl, how often, and how many pages to fetch from each site. We don't accept payment to crawl a site more frequently, and we keep the search side of our business separate from our revenue-generating AdWords service.

Google's crawl process begins with a list of web page URLs. As the Googlebot browses these websites it detects links on each page and adds them to its list of pages to crawl. The Googlebot makes a copy of each of the pages it crawls in order to compile a massive index of all the words it sees. That list also indicates where each word occurs on each page.

When a user enters a query, our machines search the index for matching pages and return the most relevant results to the user. Relevancy is determined by over 200 factors, one of which is the 'PageRank' for a given page. PageRank is the measure of the 'importance' of a page based on the incoming links from other pages. In simple terms, each web page linking to Web page XYZ adds to Web page XYZ's PageRank.



## What's new in Google web search?

While the fundamentals of web search have largely remained constant, Google is constantly working to improve its search results.

What's different from web search, say, five years ago? Well, for one it's a lot faster.

In addition, compared to five years ago our crawl and index systems are much more intelligent. For example, we now browse web pages continuously, and schedule visits to each page in a smarter way so as to maximise freshness. This more efficient approach takes into account the fact that a newspaper's online site, for example, needs frequent crawling whereas a static website updated once a month does not. In fact, we are also letting webmasters control how frequently we crawl their sites through our webmaster tools. Overall this results in a fresher and more comprehensive index.

Although web search today is faster and more efficient than ever, the key factors determining a website's visibility in the Google search results have been a priority since the day our search engine made its debut:

Can Google find the site? (page 5)

Can Google index the site? (page 6)

Does the site have unique and useful content? (page 11)



## Can Google find your site?

Inclusion in Google's search results is free and easy; you don't even need to submit your site to Google. In fact, the vast majority of sites listed in our results aren't manually submitted for inclusion, but found and added automatically when the Googlebot crawls the web.

Although Google crawls billions of pages, it's inevitable that some sites will be missed. When the Googlebot misses a site, it's frequently for one of the following reasons:

- the site isn't well connected through multiple links to other sites on the web;
- the site launched after Google's most recent crawl was completed;
- the site was temporarily unavailable when we tried to crawl it or we received an error when we tried to crawl it.

Using Google webmaster tools such as Sitemaps, you can determine whether your site is currently included in Google's index and whether we received errors when we tried to crawl it increased (see page 14). You can also use these tools to add your URL to Google's index manually, or provide Google with a Sitemap that gives us greater insight into your content. This helps us find new sections and content from your site.

# Can Google index your site?

Occasionally, webmasters will discover that their sites are not appearing in search results. The issue could be one of 'indexability' – whether or not the Googlebot can make a copy of a web page for inclusion in our search results.

### Structure and Content

A common reason for non-inclusion in search results is tied to the structure and content of the web page. For example, a page that requires a user to fill out a form may not be indexable by Google. Nor can a page using 'dynamic content' (Flash, JavaScript, frames or dynamically generated URLs) always be easily indexed by search engines. If you are wondering whether this might be your site's problem, try viewing the site in a text browser like Lynx or in a browser with images, Javascript, and Flash turned off, which will signal whether all of your content is accessible.

If your site uses a lot of images, ensure that you describe the important content of each image in the text. Not only does this allow search engines to index the image correctly, it also makes the image accessible to visually impaired users. You can also make use of alt text for the image and use descriptive filenames, as shown in this example (which is an image of a logo for a company called 'Buffy's House of Pies'):

```
<img src="buffyshouseofpies.jpg" alt="Welcome to
Buffy's House of Pies!">
```

### URLs

An additional hurdle could be the URL itself. If there are session IDs or several parameters in the URL, or if the URL redirects a number of times, Google may not be able to index the page.

### Server and Network

Server or network issues may prevent us from accessing certain pages of your site. By using tools available at Google's Webmaster Central, publishers can see a list of their pages the Googlebot cannot access. To learn more about Webmaster Central, see page 13.

### Robots Exclusion Protocol

Occasionally pages will be blocked by the Robots Exclusion Protocol, a technical standard that allows web publishers to 'tell' search engines not to index their site's content (see page 7). If your website isn't appearing in Google search results you should check to ensure that *robots.txt* or a meta tag isn't blocking access to our crawlers.



## *Controlling what Google indexes*

Every web publisher has a different goal for what he or she is trying to achieve on the Internet. Some newspaper publishers, for example, have chosen to provide free access to their recent articles, while opting for a premium, paid service for access to archives. Some want visibility on all of a search engine's properties (for example, Google Mobile, Google Images, etc.), while others only want to appear in web search results.

Search engines want to respect publishers' wishes – after all, it's their content. But we aren't mind readers, so it's vital that webmasters tell us how they want their content indexed. This can be done via the Robots Exclusion Protocol, a well-established technical specification that tells search engines which site or parts of a site should not be searchable, and which parts should remain visible in the search results.

### Robots.txt: site-wide control

The core of the Robots Exclusion Protocol is a simple text file called *robots.txt* that has been an industry standard for many years. With *robots.txt* you can control access at multiple levels, from the entire site to individual directories, pages of a specific type, or even individual pages.

> There are some pages on my site that I don't want in Google's index. How do I keep them from appearing in Google's search results?

In general, most site owners *want* the Googlebot to access their site so that their web pages can be found by users searching on Google. However, you may have pages that you don't want indexed: for example, internal logs or news articles that require payment to access.

You can exclude pages from Google's index by creating a *robots.txt* file and placing it in the root directory on your web server. The *robots.txt* file lists the pages that search engines shouldn't index. Creating a *robots.txt* file is straightforward and gives publishers a sophisticated level of control over how search engines access their websites.

For example, if a webmaster wants to prevent indexing of his internal logs the *robots.txt* file should contain:

`User-Agent: Googlebot` – The User-Agent line specifies that the next section is a set of instructions just for the Googlebot.

`Disallow: /logs/` – The Disallow line tells the Googlebot not to access files in the logs sub-directory of your site.

The site owner has specified that none of the pages in the logs directory should show up in Google's search results.

All major search engines will read and obey the instructions you put in *robots.txt*, and you can specify different rules for different search engines if you wish.

### Meta tags: fine-grain control

In addition to the *robots.txt* file – which allows you to specify instructions concisely for a large number of files on your web site – you can use the robots meta tag for fine-grain control over individual pages on your site. To implement this, simply add specific meta tags to an HTML page to control how that page is indexed. Together, *robots.txt* and meta tags give you the flexibility to express complex access policies relatively easily.

I have a particular news article on my site that is only accessible to registered users. How do I keep this out of Google's search results?

To do this, simply add the NOINDEX meta tag to the first <head> section of the article. It should look something like this:

```
<html>
<head>
<meta name="googlebot" content="noindex">
[...]
```

This stops Google from indexing this file.

However, it's worth keeping in mind that in some cases you may want Google to index these types of pages – for example, an archive newspaper article that viewers can pay to read online. While this type of "premium" content won't appear in Google's search results, certain Google services like News Archive Search will include the article in their indexes, with payment information clearly displayed to users.



## *Robots.txt vs. meta tags*

In general, *robots.txt* is a good way to provide site-wide control, while meta tags give fine-grain control over individual files. Meta tags are particularly useful if you have permission to edit individual files but not the entire site. Meta tags also allow you to specify complex access-control policies on a page-by-page basis.

Sometimes either of the two tools can solve the same problem:

How can I make sure the text of a page is indexed, but not the images?

One option would be to block access to images by file extension across your site using *robots.txt.* The following lines in a *robots.txt* file tell Google not to index any files ending in **.jpg* or **.jpeg:*

```
User-agent: Googlebot
Disallow: /*.jpg$
Disallow: /*.jpeg$
```

Alternatively, if your Content Management System (CMS) stores images in a separate directory, you can exclude that entire directory. If your images are in a directory called /images you can exclude that directory from all search engines using:

```
User-agent: *
Disallow: /images/
```

Another option would be to add a NOINDEX tag to each file that includes an image.

All these approaches will keep your images from being indexed; the only question is how extensive you would like this image exclusion to be.

## *Controlling caching and snippets*

Search results usually show a cached page link and a snippet. Here, for example, is one of the first results we see when we search for 'Mallard duck':


***Snippet*** – an extract of text from the web page

***Cached link*** – this link takes users to a copy of the page stored on Google's servers

Why have a snippet? Users are more likely to visit a website if the search results show a snippet from that site. This is because snippets make it much easier for users to see the relevance of the result to their query. If a user isn't able to make this determination quickly, he or she usually moves on to the next search result.

Why have a cached link? The cached link is useful in a number of cases, such as when sites become temporarily unavailable; when news sites get overloaded in the aftermath of a major event; or when sites are accidentally deleted. Another advantage is that Google's cached copy highlights the words employed by the user in their search, allowing a quick evaluation of the relevance of the page.

Most web publishers want Google to display both the snippet and the cached link. However, there are some cases where a site owner might want to disable one or both of these:

> My newspaper content changes several times a day. It doesn't seem like the Googlebot is indexing that content as quickly as we are updating it, and the cached link is pointing to a page that is no longer up-to-date. How can I prevent Google from creating a cached link?

The news site owner can prevent this cached link from appearing in search results by adding the NOARCHIVE tag to its page:

```
<META NAME="GOOGLEBOT" CONTENT="NOARCHIVE">
```

Similarly, you can tell Google not to display a snippet for a page via the NOSNIPPET tag:

```
<META NAME="GOOGLEBOT" CONTENT="NOSNIPPET">
```

**Note:** Adding NOSNIPPET also has the effect of preventing a cache link from being shown, so if you specify NOSNIPPET, you automatically get NOARCHIVE as well.



## Does your site have unique and useful content?

Once the site is discoverable and indexable, the final question to ask is whether the content of the web pages is unique and useful.

First look at your text as a whole. Are your title and text links descriptive? Does your copy flow naturally and in a clear and intuitive manner?

Just as a chapter in a book is organised around specific areas and themes, so each web page should be focused on a specific area or topic. Keywords and phrases emerge naturally from this type of copy, and users are far more likely to stay on a web page that provides relevant content and links.

Make sure, however, that the phrases you write include the phrases that visitors will likely search for. For instance, if your site is for an MG enthusiast club, make sure the words 'MG' and 'cars' actually appear in the copy, rather than only terms like 'British automobiles'.

## Increasing visibility: best practices

Site owners often ask us about the best ways to increase the visibility and ranking of their sites in our search results. Our simple answer is, 'Think like a user, because that's how we try to think.'

What does this mean in practice? Above all, make sure to give visitors the information they are looking for, as relevance is what will drive traffic to your site and help you retain it.

Many site owners fixate on how well their respective web pages rank. But ranking is determined by over 200 criteria in addition to PageRank. It's much better to spend your time focusing on the quality of your content and its accessibility than trying to find ways to 'game' a search engine's algorithm. If a site doesn't meet our quality guidelines, it may be blocked from the index.

### What to do:

1. Create relevant, eye-catching content: visitors will arrive at your pages via various links, so make sure each page will grab their attention.
2. Involve users: can you add a comments section or blog to your website? Building a community helps drive regular usage of your site. Getting your visitors involved helps boost visibility and user fidelity.
3. Monitor your site: use Webmaster Central (see page 13) to see what queries are pushing visitors to your site, or to track ranking changes in search results in relation to larger site changes.
4. Aim for high-quality, inbound links.
5. Provide clear text links: place text links appropriately on your site and make sure they include terms that describe the topic.

### What to avoid:

1. Don't fill your page with lists of keywords.
2. Don't attempt to 'cloak' pages by writing text that can be seen by search engines but not by users.
3. Don't put up 'crawler only' pages by setting up pages or links whose sole purpose is to fool search engines.
4. Don't use images to display important names, content or links – search engines can't 'read' images.
5. Don't create multiple copies of a page under different URLs with the intent of misleading search engines.

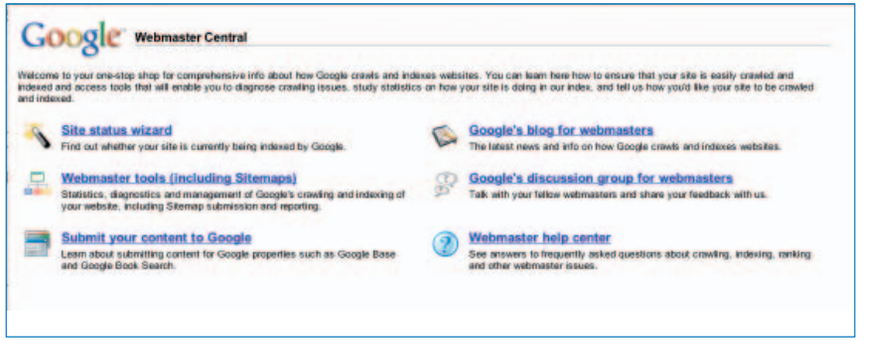
When in doubt, consult our webmaster guidelines, available at:
***google.com/webmasters/guidelines.html***



## Webmaster Central

As a company aiming to provide the most relevant and useful search results on the web, we strive to provide scalable and equitable support for all webmasters and all websites, no matter how large or small. That's why we've created Webmaster Central, located at ***google.com/webmasters***.

Webmaster Central is a great resource for all web publishers. It comprehensively answers crawling, indexing, and ranking questions; provides and avenue for feedback and issues; and offers diagnostic tools that help webmasters debug potential crawling problems.




Here's a taste of what you can find at Webmaster Central.

- Diagnose potential problems in accessing pages and offer solutions
- Request removal of specific pages from our index
- Ensure your *robots.txt* file is allowing and blocking the pages you expect
- See query and page statistics related to your website:
- Query statistics: Determine which search queries bring your site the most visitors, and what topics could your site expand upon to capture more traffic?
- Page Analysis: See your web page as Google sees it. View the most common words on your site, inbound links to the site, and how others describe your site when they link to it.
- Crawl rate: See how frequently your site is being crawled by the Googlebot and tell Google whether it should be crawling faster or slower.

### *Sitemaps*

Webmaster Central also offers publishers Sitemaps for web search, mobile, and news results.

Sitemaps is a protocol we support with other search engines to help webmasters give us more information about their web pages. Sitemaps complements existing standard web crawl mechanisms and webmasters can use it to tell Google about the pages of their site in order to improve the crawling and visibility of their pages in Google search results.

In addition to Web Search Sitemaps, we also offer Google Mobile Sitemaps, which enables publishers to submit URLs that serve content for mobile devices into our mobile index.

And for those publishers whose news site is included in Google News, News Sitemaps can help provide statistics about the publisher's articles, from queries to frequency of appearance. Used in conjunction with Webmaster Central diagnostic tools, News Sitemaps can also provide error reports that help explain any problems Google experiences when crawling or extracting news articles from a publisher's site. In addition, a publisher can submit a News Sitemap containing URLs that it would like considered for inclusion in Google News. News Sitemaps, unlike Web and Mobile Sitemaps, is currently only available in English, although we hope to soon make it available in other languages as well.



## Frequently Asked Questions

Why can't you do one-on-one support for my website?
By some estimates, there are 100 million sites on the web. Each of those websites is important to us, because without them – no matter how big or small – our index would be less comprehensive, and ultimately less useful to our users.

Webmaster Central is a great source of support for all types of websites. We post and answer publishers' questions so that everyone can benefit from the information. At Webmaster Central you can also find a friendly and useful community of webmasters who can share tips and help with troubleshooting.

Do the ads you show influence your rankings? Are your ad listings and search results totally separate?
Ad and search rankings are not related in any way, and in fact we have entirely separate teams to work on them so that there is no interference. We believe that the objectivity of our search results is crucial to providing the best experience for users.

How do I add a site to Google's search index?
Inclusion in Google's search results is free and easy and does not require a manual submission of the site to Google. Google is a fully automated search engine; it crawls the web on a regular basis and finds sites to add to our index. In fact, the vast majority of sites listed in our results aren't manually submitted for inclusion, but found and added automatically when our spiders crawl the web.

In addition, Google Webmaster Tools (at Webmaster Central) provide an easy way for webmasters to submit a sitemap or their URLs to the Google index and get detailed reports about the visibility of their pages on Google. With Google Webmaster Tools, site owners can automatically keep Google informed of all current pages and of any updates made to those pages.

How long, on average, does it take for Google to find a newly created website, and how frequently does Google crawl the web in general?
There is no set amount of time it takes for Google to find a new site.

The Googlebot regularly crawls the web to rebuild our index. By using Webmaster Central, a webmaster can see how frequently his site is being crawled by the Googlebot and tell Google whether it should be crawling faster or slower.

What if I want my website to appear on web search results, but not on separate services like Google News or Google Image Search?
Google always allows web publishers to opt out of its services, and a publisher can contact the support team of a particular product to do so.

As discussed earlier in this booklet, the Robots Exclusion Protocol can be used to block indexing of both images and web pages. The 'URL removal' feature in Webmaster Central can also be used for this purpose and covers web search and image search.

In addition, because the Googlebot relies on several different bots, you can target what you block:

- Googlebot: crawl pages from our web index and our news index
- Googlebot-Mobile: crawls pages for our mobile index
- Googlebot-Image: crawls pages for our image index
- Mediapartners-Google: crawls pages to determine AdSense content. We only use this bot to crawl your site if you show AdSense ads on your site.
- Adsbot-Google: crawls pages to measure AdWords landing page quality. We only use this bot if you use Google AdWords to advertise your site.

For instance, to block Googlebot entirely, you can use the following syntax:

```
User-agent: Googlebot
Disallow: /
```

### Can I choose what text I want specified as a snippet?

No. This isn't a good idea, both from the perspective of the user and the content creator. We choose a snippet of text from the site that shows the searcher's query in context, which in turn demonstrates the relevance of the result.

Studies show that users are more likely to visit a website if the search results show the snippet. This is because snippets make it much easier for users to see why the result is relevant to their query. If a user can't make this determination quickly, he or she usually moves on to the next search result.

Web publishers can include a meta tag in their pages in order to provide Google with additional input in cases where we aren't able to generate a useful snippet from the content on the page algorithmically. To do this, simply add the following to the <head> section of the page:

```
<meta name="description" content="Why doesn't Anya
like bunnies? We're about to find out.">
```

Any web publisher who doesn't want a snippet generated from their pages can use the NOSNIPPET tag, as follows:

```
<meta name="robots" content="nosnippet">
```

Finally, we sometimes use a site's description from the Open Directory Project for the search result snippet. If you don't want this description to be used, simply add the following meta tag:

```
<meta name="robots" content="noodp">
```



### Breaking news articles on my site only appear for a few hours before being updated and moved to a standard articles section. I want the full article to appear in Google's index, but not these breaking news stories.

One option is to put all the breaking news articles in one directory, and use *robots.txt* to disallow the Googlebot access to that directory.

Another option is to add the NOFOLLOW tag to the <HEAD> section of the html of your breaking news section. This tells the Googlebot not to follow any links it finds on that page. Keep in mind, though, that NOFOLLOW only stops the Googlebot from following links from one page to another. If another web page links to that article, Google can still find and index *promnight.html* when it indexes.

### If I have multiple domain names and run the same content off these different domains will I be kicked out of your search results?

Although some publishers may try to fool search engines by duplicating content and running mirror sites, there is also legitimate content that may be duplicated for good reasons. Google doesn't want to penalise those sites. For example, we don't treat similar content expressed in different languages (say, English on one site and French in another) as duplicate content.

Having the same content on multiple websites (e.g. article syndication) won't necessarily result in one or some of those sites being removed from the search results entirely. However, keep in mind that each instance of the article is likely to appear lower in the rankings because it only has a fraction of the incoming links that a single copy would. In general, a single copy of an article will rank higher and therefore be seen by more users than will multiple copies of the same content.

In addition, in order to ensure search quality, Google doesn't include multiple copies of a page in our search results. Rather, we often choose just one version of the page to show. However, webmasters can indicate to Google their preferred version by using *robots.txt* or a meta tag to block any copies they don't want showing up in our search results.

### Why is my site being blocked from the Google index?

First, your site may not be blocked. There are many reasons why a site may not appear in our search results (see pages 5-11).

If your site isn't presenting any hurdles for discovery or indexation, then it could be that your site is blocked. Sites may be blocked from our index because they do not meet the quality standards outlined in our Webmaster Guidelines (available at Webmaster Central). This most often happens when a website is using unfair methods to try to appear higher in the search rankings. Common guideline violations include cloaking (writing text in such a way that it can be seen by search engines but not by users) or setting up pages/links with the sole purpose of fooling search engines and manipulating search engine results.

When webmasters suspect that their sites violate our quality guidelines, they can modify their site to meet these guidelines, then click the "request re-inclusion" link within our Webmaster Tools interface to ask us to re-evaluate the site.

# Glossary

***Cache link***
A snapshot of how a page appeared the last time Google visited it. A cached copy allows users to view a page even when the live version is unavailable, although the content may differ slightly. To view a cached copy, click on the 'cached' link that appears underneath a search result.

***Cloaking***
Showing search engines different content than what you show users.

***CMS (Content Management System)***
A software system used to manage content from computer, image, and audio files to web content.

***Crawler***
Software used to discover and index URLs on the web or an intranet.

***Crawling***
The process used by search engines to collect pages from the web.

***Dynamic content***
Content such as images, animations, or videos which rely on Flash, JavaScript, frames, or dynamically generated URLs.

***File extension***
The name of a computer file (.doc, .txt, .pdf, etc.) often used to indicate the type of data stored in the file.

***HTML (Hypertext Markup Language)***
A mark-up language used on the web to structure text.

***To index***
The process of having your site's content added to a search engine.

***Keyword***
A word that is entered into the search box of a search engine. The search engine then looks for pages that include the word or phrase.

***Meta tags***
A tag in the HTML that describes the content of a web page. Meta tags can be used to control indexing of individual pages in a website.

***Mirror site***
A duplicate web page; sometimes used to fool search engines and try to optimise indexation and web rankings of a website.

### *Page Rank*
A Google feature that helps determine the rank of a site in our search results. PageRank relies on the uniquely democratic nature of the web by using its vast link structure as an indicator of an individual page's value. Important, high-quality sites receive a higher PageRank, which Google remembers each time it conducts a search. Google combines PageRank with sophisticated text-matching techniques to find pages that are both important and relevant to your searches.

### *Robots Exclusion protocol*
A technical specification that tells search engines which site or parts of a site should be non-searchable, and which parts should remain visible in the search results.

### *Robots.txt*
A text file that allows a web publisher to control access to their site at multiple levels, from the entire site to individual directories, pages of a specific type, or even individual pages. This file tells crawlers which directories can or cannot be crawled.

### *Root directory*
The top or core directory in a computer file system.

### *URL (Uniform Resource Locator)*
The address of a website on the Internet, consisting of the access protocol (http), domain name (www.google.com), and in some cases the location of another file (www.google.com/webmaster).

For more info about Webmaster Central please visit:

**google.com/webmasters/**

Google™