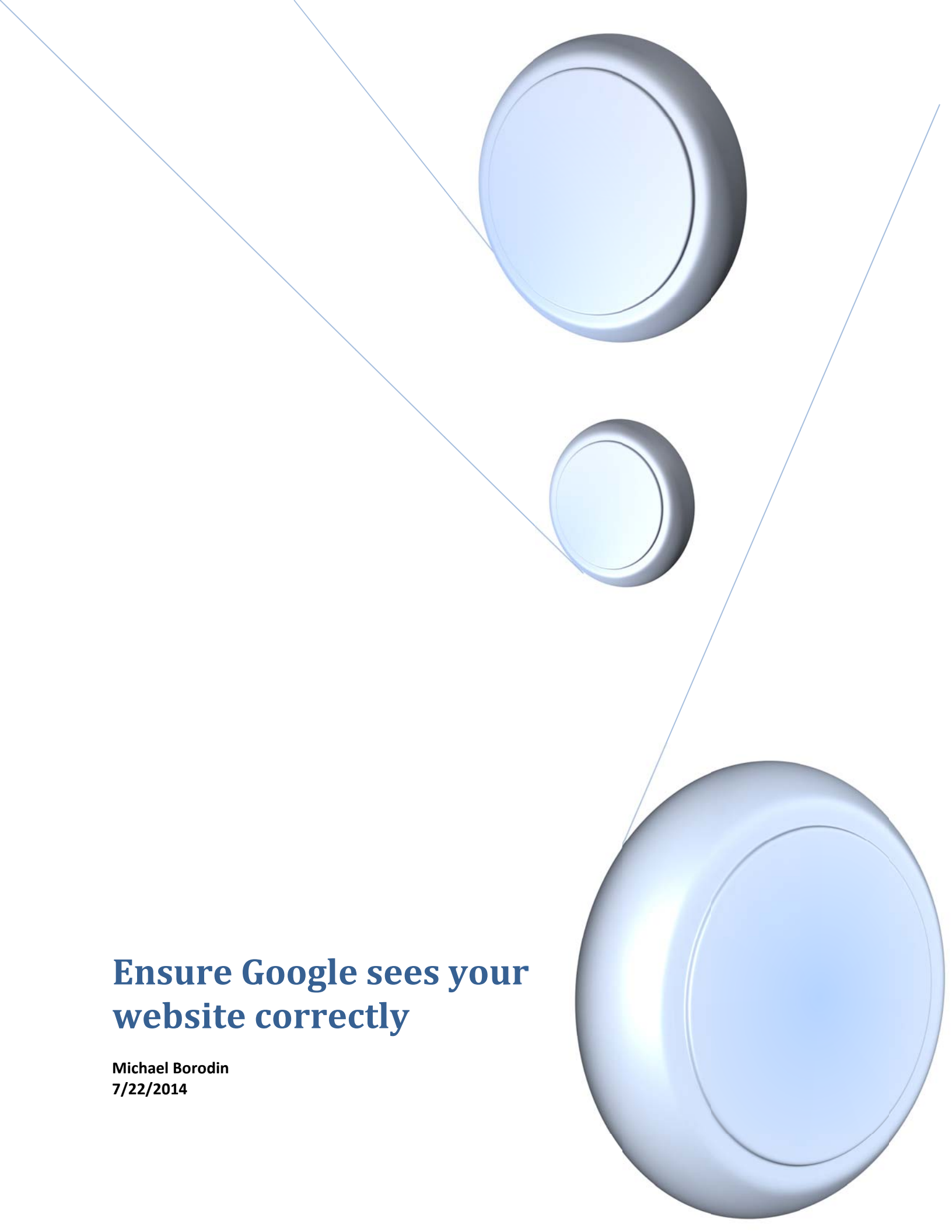

Ensure Google sees your website correctly

Michael Borodin
7/22/2014

Contents

Purpose

For the Client to have the best online experience for their company, we will need to ensure that Google sees the website correctly. Doing this will increase the ranking of the website. In order to this the Google search engine crawlers need to be able to crawl the Client's website.

Controlling Crawling and Indexing

Introduction

Search engines generally go through two main stages to make content available for users in search results: **crawling** and **indexing**. Crawling is when search engine crawlers access publicly available webpage. In general, this involves looking at the webpages and following the links on those pages, just as a human user would. Indexing involves gathering together information about a page so that it can be made available ("served") through search results.

The distinction between crawling and indexing is critical. Confusing on the point is common and leads to webpages appearing or not appearing in search results. Note that a page may be crawled but not index; and, in rare cases, it may be indexed even if it hasn't been crawled. Additionally, in order to properly prevent indexing of a page, you must allow crawling or attempted crawling of the URL.

In some situations, you may not want to allow crawlers to access areas of a server. This could be the case if accessing those pages uses the limited server resources, or if problems with the URL and linking structure would create an infinite number of URLs if all of them were to be followed.

Controlling Crawling

The robots.txt file is a text file that allows you to specify how you would like your site to be crawled. Before crawling a website, crawlers will generally request the robots.txt file from the server. Within the robots.txt file, you can include sections for specific (or all) crawlers with instructions ("directives") that let them know which parts can or cannot be crawled.

Location of the robots.txt file

The robots.txt file must be located on the root of the website host that it should be valid for. For instance, in order to control crawling on all URLs below http://www.example.com/, the robots.txt file must be located at http://www.example.com/robots.txt. A robots.txt file can be placed on subdomains (like http://website.example.com/robots.txt) or non-standard ports (http://example.com:8181/robots.txt), but it cannot be placed in a subdirectory (http://example.com/pages/robots.txt).

File Location & range validity

The robots.txt file must be in the top-level directory of the host, accessible though the appropriate protocol and port number. Generally accepted protocols for robots.txt (and crawling of websites) are "http" and "https". On http and https, the robots.txt file is fetched using a HTTP non-conditional GET request.

Google-specific: Google also accepts and follows robots.txt files for FTP sites. FTP-based robots.txt files are accessed via the FTP protocol, using an anonymous login.

The directives listed in the robots.txt file apply only to the host, protocol and port number where the file is hosted.

Note: the URL for the robots.txt file is - like other URLs - case-sensitive.

Content of the robots.txt file

You can use almost any text editor to create a robots.txt file. The text editor should be able to create standard ASCII or UTF-8 text files; don’t use a word processor (word processors often save files in a proprietary format and can add unexpected characters, such as curly quote, which may cause problems for crawlers).

A general robots.txt file might look like this:

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

User-agent: *
Disallow: /onlygooglebot/

Sitemap: http://www.example.com/sitemap.xml
```

Assuming the file is located at http://example.com/robots.txt, it specifies the following directives:

1. No Googlebot crawlers should crawl the folder http://example.com/nogooglebot/ and all contained URLs. The line "**User-agent: Googlebot**" starts the section with directives for Googlebots.
2. No other crawlers should crawl the folder http://example.com/onlygooglebot/ and all contained URLs. The line "**User-agent: *"** starts the section for all crawlers not otherwise specified.
3. The site's Sitemap file is located at http://example.com/sitemap.xml

Robot Instruction Meanings

User-agent:

The "User-agent" part is there to specify directions to a specific robot if need. There are two ways to use this in your fie.

If you want to tell all robots the same thing you put a "*" after the "User-agent" it would look like this

User-agent: *

This line is saying "these directions apply to all robots" if you want to tell a specific robot something (in this example Googlebot) it would look like this

User-agent: Googlebot

(This line is saying "these directions apply to just Googlebot)

Disallow

The “Disallow” part is there to tell you the robots what folders they should not look at. This means that if, for example you do not want search engines to index the photos on your site then you can place those photos into one folder and exclude it.

Let’s say that you have put all these photos in a folder called “photos”. Now you want to tell search engines not to index that folder.

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

The above two lines of text in your robots.txt file would keep robots from visiting your photos folder. The “User-agent *” part is saying “this applies to all robots”. The “Disallow: /photos” part is saying “don’t visit or index my photos folder”.

Googlebot specific instructions

The robot that Google uses to index their search engine is called Googlebot. It understands a few more instructions than other robots. The instructions it follows are well defined in the Google help pages.

In addition to the “User-name” and “Disallow” Googlebot also uses the

```
Allow:
```

The “Allow:” instructions lets you tell a robot that it is okay to see a file in a folder that has been “Disallowed” by other instruction.

```
User-agent: *
Disallow: /photos
Allow: /photos
```

This would tell Googlebot that it can visit “photos” folder.

Key Concept

- If you use a robots.txt file, make sure it is correctly written because an incorrect robots.txt file can block the bots that index your website

Samples of robots.txt files

These are some simple samples to help get started with the robots.txt handling

Allow crawling of all content

```
User-agent: *
Disallow:
```

Or

```
User-agent: *
Allow: /
```

The sample above is valid, but in fact if you want all your content to be crawled, you don't need a robots.txt file at all (and we recommend that you don't use one). If you don't have a robots.txt file, verify that your hoster returns a proper 404 "Not found" HTTP result code when the URL is requested.

Disallow crawling of the whole website

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

Keep in mind that in some situations URLs from the website may still be indexed, even if they haven’t been crawled.

Disallow crawling of certain parts of the website

```
User-agent: *
Disallow: /calendar/
Disallow: /junk/
```

Remember that you shouldn't use robots.txt to block access to private content: use proper authentication instead. URLs disallowed by the robots.txt file might still be indexed without being

crawled, and the robots.txt file can be viewed by anyone, potentially disclosing the location of your private content.

Allowing access to a single crawler

```
User-agent: Googlebot-news
Disallow:

User-agent: *
Disallow: /
```

Allowing access to all but a single crawler

```
User-agent: Unnecessarybot
Disallow: /

User-agent: *
Disallow:
```

Allow a single robot

```
User-agent: Googlebot
Disallow:

User-agent: *
Disallow: /
```

Allow all robots complete access

```
User-agent: *
Disallow:
```

Exclude all robots from the entire website

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

Controlling Indexing and Serving

Indexing can be controlled on a page-by-page basis using simple information that is sent with each page as it is crawled. For indexing Control, you can use either:

- A special Meta Tag that can be embedded in the top of HTML pages
- A special HTTP header element that can be sent with all content served by the website

Note: Keep in mind that in order for a crawler to find a meta tag or HTTP header element, the crawler must be able to crawl the page—it cannot be disallowed from crawling with the robots.txt file.

Using the Robots Meta Tag

The robots Meta Tag can be added to the top of a HTML page, in the <head> section, for instance:

```
<!DOCTYPE html>
<html><head>
<meta name="robots" value="noindex" />
…
```

In this example, robots Meta Tag are specifying that no search engines should index this particular page (noindex). The name robot applies to all search engines. If you want to block or allow a specific search engine, you specify a user-agent name in the place of robots.

Valid indexing & serving directives

Several other directives can be used to control indexing and serving with the robots Meta tag and the X-Robots-Tag. Each value represents a specific directive. The following table shows all the directives that Google honors and their meaning. Note: it is possible that these directives may not be treated the same by all other search engine crawlers. Multiple directives may be combined in a comma-separated list (see below for the handling of combined directives). These directives are not case-sensitive.

Directive	Meaning
all	There are no restrictions for indexing or serving. Note: this directive is the default value and has no effect if explicitly listed.

noindex	Do not show this page in search results and do not show a "Cached" link in search results.
nofollow	Do not follow the links on this page
none	Equivalent to noindex, nofollow
noarchive	Do not show a "Cached" link in search results.
nosnippet	Do not show a snippet in the search results for this page
noodp	Do not use metadata from the Open Directory project for titles or snippets shown for this page.
notranslate	Do not offer translation of this page in search results.
noimageindex	Do not index images on this page.
unavailable_after: [RFC-850 date/time]	Do not show this page in search results after the specified date/time. The date/time must be specified in the RFC 850 format.

Handling combined indexing and serving directives

You can create a multi-directive instruction by combining robots Meta Tag directives with commas. Here is an example of robots Meta Tag that instructs web crawlers to not index the page and to not crawl any of the links on the page:

```
<meta name="robots" content="noindex, nofollow">
```

For situations where multiple crawlers are specified along with different directives, the search engine will use the sum of the negative directives. For example:

```
<meta name="robots" content="nofollow">
<meta name="googlebot" content="noindex">
```

The page containing these Meta Tags will be interpreted as having a noindex, nofollow directive when crawled by Googlebot.

File Format

The expected file format is plain text encoded in UTF-8. The file consists of records (lines) separated by CR, CR/LF or LF.

Only valid records will be considered; all other content will be ignored. For example, if the resulting document is a HTML page, only valid text lines will be taken into account, the rest will be discarded without warning or error.

If a character encoding is used that results in characters being used which are not a subset of UTF-8, this may result in the contents of the file being parsed incorrectly.

A maximum file size may be enforced per crawler. Content which is after the maximum file size may be ignored. Google currently enforces a size limit of 500kb.

Groupings of records

Records are categorized into different types based on the type of <field> element:

- Start-of-group
- Group-member
- Non-group

All group-member records after a start-of-group record up to the next start-of-group record are treated as a group of records. They only start-of-group field element is user-agent. Multiple start-of-group lines directly after each other will follow the group-member records following the final start-of-group line. Any group-member records without a preceding start-of-group record are ignored. All non-group records are valid independently of all groups.

Valid <filed> elements, which will be individually detailed further on in this document, are:

- User-agent (start of group)
- Disallow (only valid as a group-member record)
- Allow (only valid as a group-member record)
- Sitemap (non-grouped record)

All other <field> elements may be ignored.

The start –of-group element user-agent is used to specify for which crawler the group is valid. Only one group of records is valid for a particular crawler.

Example groups:

```
user-agent: a
disallow: /c

user-agent: b
disallow: /d

user-agent: e
user-agent: f
disallow: /g
```

There are three distinct groups specified, one for “a” and one for “b” as well as one for both “e” and “f”. Each group has its own group-member record. Note the optional use of white-space (an empty line) to improve readability.

Order of precedence for user-agents

Only one group of group-member is valid for a particular crawler. The crawler must determine the correct group of records by finding the group with the most specific user-agent that still matches. All other groups of records are ignored by the crawler. The user-agent is non-case-sensitive. All non-matching text is ignored (for example, both Googlebot/1.2 and Googlebot* are equivalent to Googlebot). The order of the groups within the robots.txt file is irrelevant.

Example

Assuming the following robots.txt file:

```
user-agent: googlebot-news
(group 1)

user-agent: *
(group 2)

user-agent: googlebot
(group 3)
```

This is how the crawlers would choose the relevant group:

Name of Crawler	Record group followed	Comment
Googlebot News	(group 1)	Only the most specific group is followed, all others are ignored.
Googlebot (web)	(group 3)	
Googlebot Images	(group 3)	There is no specific googlebot-images group, so the more generic group is followed.
Googlebot News (when crawling images)	(group 1)	These images are crawled for and by Googlebot News, therefore only the Googlebot News group is followed.
Otherbot (web)	(group 2)	
Otherbot (News)	(group 2)	Even if there is an entry for a related crawler, it is only valid if it is specifically matching.

Google-supported non-group-member records

Sitemap

Supported by Google, Ask, Bing, and Yahoo; defined on sitemaps.org.

Usage:

```
sitemap: [absoluteURL]
```

[AbsoluteURL] points to a Sitemap, Sitemap Index file or equivalent URL. The URL does not have to be on the same host as the robots.txt file. Multiple sitemap entries may exist. As non-group-member records, these are not tied to any specific user-agents and may be followed by all crawlers, provided it is not disallowed.

How to examine your site the same way as Google does

Google provides a “Page Spider Tool” that uses the same methods that Googlebot does to look at a webpage. Using this tool you can make sure that the search engine spiders are seeing the web pages completely.

http://www.feedthebot.com/tools/spider/

Make sure the text is being seen and the links are being seen. These are the biggest issues and can be blocked by things like Flash or JavaScript (the things Google calls “Fancy Features”).

Examining your website using lynx

To examine your site in Lynx you must download and install it to your computer. When I went to the Lynx homepage, I found it somewhat unclear and difficult to install. The first thing I noticed when I used Lynx was that it was like an old DOS window. For those of you who were not using computers in the "early days" when DOS was used often, I can tell you it will seem a bit confusing.

Now if you were to navigate through this page in the text browser, you would find that each link works. You will also find that all the text is displayed, and that the search form works. This means that all the elements of the Google home page work also in text browsers. Congratulations Google! Your home page does not pose a problem to search engine crawlers.

If your website can be navigated through in a text browser, then search engine crawlers can navigate it as well.

A simpler way to check your pages is through a "spider simulator" which shows you an approximation of what a search engine crawler might see on you site. If you check your site on a spider simulator you will be able to detect most of the obvious problems (like text or links not being visible).

Key Concepts:

Make sure that search engine spiders are able to see your site correctly. Ensuring that your website is seen correctly by search engine spiders is vital.

Appendix: Google's website crawlers

Crawler	User-agents	User-agent in HTTP(s) request	Comments + Documentation URL
Googlebot (web)	Googlebot	Mozilla/5.0 (compatible; Googlebot/2.1; +http://www.google.com/bot.html) Alternate (rarely used): Googlebot/2.1 (+http://www.google.com/bot.html)	Generic Googlebot crawler for web-search http://www.google.com/bot.html
Googlebot News	Googlebot-News (Googlebot)	Googlebot-News	For Google News
Googlebot Images	Googlebot-Image (Googlebot)	Googlebot-Image/1.0	For Image Search
Googlebot Video	Googlebot-Video (Googlebot)	Googlebot-Video/1.0	For Video Search
Googlebot Mobile	Googlebot-Mobile (Googlebot)	[various mobile device types] (compatible; Googlebot-Mobile/2.1; +http://www.google.com/bot.html)	For Google Mobile web-search results http://www.google.com/support/mobile/bin/answer.py?answer=37425
Google Mobile AdSense	Mediapartners-Google Mediapartners (Googlebot)	[various mobile device types] (compatible; Mediapartners-Google/2.1;	

		+http://www.google.com/bot.html)	
Google AdSense	Mediapartners-Google Mediapartners (Googlebot)	Mediapartners-Google	
Google AdsBot landing page quality check	AdsBot-Google (Given the special nature of this crawler, only directives for this user-agent are followed.)	AdsBot-Google (+http://www.google.com/adsbot.html)	Only visits landing pages used in AdWords campaigns. **See http://www.google.com/adsbot.html**

Appendix: Robots Database

http://www.robotstxt.org/db.html

Appendix: Googlebot

Name	Googlebot
Cover	http://www.googlebot.com/
Details	http://www.googlebot.com/bot.html
Status	active
Description	Google's crawler
Purpose	indexing
Type	standalone
Platform	Linux
Language	c++
Availability	none
Owner Name	Google Inc.
Owner URL	http://www.google.com/
Owner Email	googlebot@google.com
Exclusion	yes
Exclusion User-Agent	googlebot
NOINDEX	yes
Host	googlebot.com
From	yes
Useragent	Googlebot/2.X (+http://www.googlebot.com/bot.html)
History	Developed by Google Inc.
Environment	commercial
ID	googlebot
Modified Date	Thu Mar 29 21:00:07 PST 2001
Modified By	googlebot@google.com