Как составить правильный robots.txt для сайта на wordpress

Что такое robots.txt и зачем он нужен?

Файл **robots.txt** - это служебный файл, в котором можно указать при помощи специальных директив, какие разделы сайта индексировать, а какие нет.
Директивы можно указывать, как для всех роботов сразу, так и для каждой поисковой системы в отдельности.

Разделы, закрытые от индексации, не попадут в индекс поисковых систем.

Что такое Индекс?

Это база данных поисковой системы, в которой она хранит набор встречающихся на интернет-страницах слов и словосочетаний.

Эта информация соотнесена с адресами тех веб-страниц, на которых она встречаются, и постоянно пополняется новой информацией, собираемой роботом-пауком поисковой системы.

Для того, чтобы сайт появлялся в выдаче поисковой системы по определенным запросам, он должен быть занесен в индекс этой поисковой системы.

Вообщем, **robots.txt** – это очень полезный и нужный любому сайту файл.

Общая для всех сайтов часть файла:

Итак, создаем текстовый документ с названием robots.txt и пишем в него следующее:

```
User-agent: *
Allow: /wp-content/uploads/
Disallow: /cgi-bin/
Disallow: /wp-admin/
Disallow: /wp-includes/
Disallow: /wp-content/
Disallow: /wp-*
Disallow: */trackback
Disallow: */comments
Disallow: */comment-*
Disallow: */feed
Disallow: /?s=
Disallow: /xmlrpc.php
Host: www.yourdomain.ru

Sitemap: http://yourdomain.ru/sitemap.xml

Google Image:
User-agent: Googlebot-Image
Allow: /*
```

Internet Archiver Wayback Machine

User-agent: ia_archiver

Disallow: /

Пояснения:

User-agent: * - Директива всем роботам

Если нужно, чтобы эти правила работали только для одного, конкретного робота, то вместо * указываем его имя (User-agent: Yandex, User-agent: Googlebot и т.д.).

Allow: /wp-content/uploads/ - Разрешение индексировать папку uploads (а значит и расположенные в ней картинки).

Правило Яндекса для robots.txt гласит:

"Если для какого-то раздела сайта подходят несколько директив, то выбирается первая в порядке появления ее в выбранном блоке."

Поэтому для Яндекса директиву Allow мы указываем первой в списке.
Google понимает Allow и вверху и внизу секции.

Disallow: /cgi-bin/

Disallow: /wp-admin/

Disallow: /wp-includes/

Disallow: /wp-content/

Здесь мы запрещаем индексировать указанные папки (и все, что в них находится), кроме папки uploads, которая находится внутри папки wp-content и которую мы разрешили индексировать выше.

Disallow: /wp-* Запрещает индексацию всех файлов движка, начинающихся с wp- (т.е. всех файлов, расположенных в корневой папке).

Disallow: */trackback

Disallow: */comments

Disallow: */comment-*

Disallow: */feed

Запрещаем индексацию комментариев, трекбеков и фида.

Спецсимвол * означает любую (в том числе пустую) последовательность символов.

Disallow: /?s= Запрещаем индексацию результатов поиска.

Host: www.glavnoye-zerkalo.ru - Директива Host понимается только Яндексом и не понимается Гуглом.

В ней указывается главное зеркало сайта, в случае, если у вашего сайта есть зеркала. В поиске будет участвовать только главное зеркало.

Директиву Host лучше написать сразу после директив Disallow (для тех роботов, которые не полностью следуют стандарту при обработке robots.txt).

Для Яндекса директива Host является межсекционной, поэтому будет найдена роботом не зависимо от того, где она указана в файле.

Важно: Директива Host в файле robots.txt может быть только одна.
В случае указания нескольких директив, использоваться будет первая.

Sitemap: http://mysite.ru/sitemaps.xml - Указываем путь к файлу sitemaps.xml

Для Яндекса и Google Sitemap - это межсекционная директива, но лучше написать ее в конце через пустую строку. Так она будет работать для всех роботов сразу.

Google Image
User-agent: Googlebot-Image
Allow: /*

Разрешает Googlebot'у индексировать наши картинки.

Предупреждая вопросы, которые обязательно возникнут, скажу:

Да, эта директива отчасти повторяет директиву Allow для папки uploads.
Но, картинки ведь могут находится не только в этой папке .

Internet Archiver Wayback Machine
User-agent: ia_archiver
Disallow: /

Полностью запрещаем роботу веб архива индексацию нашего сайта.
Это предупредительная мера, которая защитит от массового парсинга контента нашего сайта через веб архив.

Итак, мы рассмотрели стандартную часть файла robots.txt, которая подойдет для любого сайта на wordpress.

Но, нам нужно спрятать от роботов-индексаторов еще некоторые разделы сайта, в частности, те, которые создают дублированный контент - разного рода архивы.

Индивидуальные настройки:

Страницы архивов на разных сайтах имеют различные адреса, в зависимости от того, как формируются URL на сайте, включены ли ЧПУ или нет.

1. Как определить адреса архивов вашего сайта?

Для этого нужно открыть архив любого месяца и посмотреть, как выглядит адрес страницы архива.

Он может выглядеть, например, так: http://sait.ru/archives/date/post-1.

В этом случае, выделяем общую для всех архивов по дате часть адреса:

/archives/date/

Соответственно, в файле robots.txt указываем:

Disallow: /archives/date/*

Будьте внимательны! - у вас архивы могут иметь другие адреса.

Например, архив года может иметь такой адрес: http://sait.ru/2012

Тогда закрывать нужно эту часть - **/2012**

Напоминаю, что спецсимвол * означает любую последовательность символов, т.е. все, что находится в адресе далее.

2. Точно так же определяем адреса архивов тегов и архивов автора.

И закрываем их в robots.txt.

Disallow: /archives/tag/*
Disallow: /archives/author/*

Рекомендую архивы тегов закрыть примерно на полгода-год, (если у вас молодой сайт) пока он стабильно пропишется в поисковой выдаче.

После этого срока архивы тегов стоит открыть для индексации, так как по наблюдениям, на страницы тегов по поисковым запросам приходит значительно больше посетителей, чем на страницы постов, к которым эти теги созданы.

Но, не добавляйте к постам много тегов, иначе вашему сайту будут грозить санкции за дублированный контент.

Один-два (редко три) тега к одному посту вполне достаточно.

3. Можно, также, закрыть индексацию всех главных страниц, кроме первой.

Обычно, их адреса выглядят так: http://sait.ru/page/2, http://sait.ru/page/3 и т. д., но лучше проверить.

Перейдите по ссылкам навигации внизу Главной страницы на вторую страницу и посмотрите, как выглядит ее адрес в адресной строке.

Закрываем эти страницы:

Disallow: /page/*

4. Иногда, на сайте требуется закрыть от индексации еще какие то страницы или папки.

В этом случае, действуете аналогично - открываете в браузере нужную вам страницу и смотрите ее адрес. А дальше закрываете его в robots.txt.

Учтите, что если вы закрываете, например, папку "**book**", то автоматически закроются и все файлы, расположенные в этой папке.

Если закрыть страницу, в URL'е которой есть, например, "**news**" так: ***/news/**,
то закроются и страницы **/news/post-1** и **/category/news/**.

robots.txt полностью:

User-agent: *
Allow: /wp-content/uploads/
Disallow: /cgi-bin/
Disallow: /wp-admin/
Disallow: /wp-includes/
Disallow: /wp-content/
Disallow: /wp-*
Disallow: */trackback
Disallow: */comments
Disallow: */comment-*
Disallow: */feed
Disallow: /?s=
Disallow: /xmlrpc.php
Disallow: /archives/date/*
Disallow: /archives/tag/*
Disallow: /archives/author/*
Disallow: /page/*
Host: www.yourdomain.ru

Sitemap: http://yourdomain.ru/sitemap.xml

Google Image:
User-agent: Googlebot-Image
Allow: /*

Internet Archiver Wayback Machine
User-agent: ia_archiver
Disallow: /

Справка:

Подробности составления robots.txt можно изучить на странице помощи Яндекса - http://help.yandex.ru/webmaster/?id=996567

Проверить правильность составления файла можно в http://webmaster.yandex - Настройка индексирования - Анализ robots.txt

В Google robots.txt можно проверить по этой ссылке:

https://www.google.com/webmasters/tools/crawl-access?hl=ru&siteUrl=http://site.ru/

Проверяемый сайт должен быть добавлен в панель веб-мастера.

P.S. Для Яндекса и Google правила составления robots.txt немного различаются.

Толкование правил составления robots.txt можно прочитать здесь - http://robotstxt.org.ru

К сожалению, проверить можно только синтаксис.
Правильно ли вы закрыли от индексации разделы сайта, покажет только время .

Ну вот и все - файл robots.txt готов, осталось только загрузить его в корневой каталог нашего сайта.

Напоминаю, что корневой каталог это папка в которой находится файл config.php.

И последнее - все, сделанные вами изменения в robots.txt, будут заметны на сайте только спустя несколько месяцев.

В тему:

Однажды видела сайт на wordpress, на котором не было файла robots.txt.
Этот сайт некоторое время простоял пустым - т.е. с одной стандартной записью, которая по умолчанию присутствует в wordpress.

Представьте себе, как «обрадовался» владелец сайта, когда обнаружил, что Яндекс проиндексировал 2 страницы с контентом и больше тысячи страниц самого движка.

На этой веселой ноте заканчиваю

Ксана

http://prodengiblog.ru

Авторские шаблоны для wordpress,

Секреты разработки и управления wordpress-сайтом

Связаться со мной - http://prodengiblog.ru/kontakty

Подписаться на обновления блога по RSS - http://feeds.feedburner.com/ProDengiBlog

Подписаться на обновления блога по email - http://prodengiblog.ru/kurs

Следите за мной в Твиттере - http://twitter.com/Ksanea

Присоединиться ВКонтакте - http://vk.com/wordpress_themes

Присоединиться в Facebook - http://www.facebook.com/pages/prodengiblogru/236880776336554

Смотреть канал на youtube - http://www.youtube.com/user/Ksanea