ЛІГАбізнесінформ
Українська мережа ділової інформації

robots.txt

стандарт, расширения, аспекты применения

www.liga.net

Поисковая оптимизация и продвижение сайтов в Интернете

Владимир Чернышов

ЛІГАбізнесінформ
Українська мережа ділової інформації

Что такое robots.txt

www.liga.net

Файл robots.txt – основной инструмент, с помощью которого вебмастер может управлять индексацией сайта роботами поисковых систем.

Основная функция файла – запрещающая, с помощью нескольких, относительно простых правил, записанных в обычном текстовом файле можно запретить индексацию страницы или группы страниц на сайте.

ЛІГАбізнесінформ
Українська мережа ділової інформації

Зачем запрещать индексацию?

www.liga.net

Как правило, запрещают индексацию неинформативных и служебных страниц

- Версии для печати
- Формы регистрации, аутентификации и т.п.
- Профили пользователей на форумах
- Корзина товаров в интернет-магазинах
- Варианты сортировки списков ссылок
- Адреса с идентификаторами сессий
- Адреса с метками

История протокола исключений

- В начале 90-х годов участились случаи, когда роботы вызывали сбои в работе веб-серверов из-за слишком высокой нагрузки при большой частоте запросов к серверу.
- Документ, описывающий протокол исключений был впервые представлен 30 июня 1994 года в специализированной рассылке.
- За 12 прошедших лет протокол так и не стал стандартом «де юре», хотя является стандартом «де факто»

Выбор имени файла

- Имя файла должно соответствовать основным критериям к именованиям файлов в большинстве операционных систем.
- Имя файла не должно требовать дополнительных настроек веб-сервера.
- Имя файла должно указывать на его предназначение и быть легко запоминаемым.
- Вероятность совпадения имени с существующими файлами должна быть минимальной.

Формат файла robots.txt

- Файл robots.txt должен находиться в корневой директории домена или поддомена
- Имя файла регистрозависимое и должно состоять только из строчных (lower-case) символов
- Записи (секции) в файле разделяются пустыми строками
- Перевод строки может быть в формате любой операционной системы, CR LF, LF или CR
- Запись состоит из одной или нескольких строк с User-agent, за которыми следуют одна или несколько строк с Disallow

Пример файла robots.txt

```
# Start
User-agent: Googlebot
User-agent: StackRambler
Disallow: /dir
Disallow: /file.htm

User-agent: *
Disallow:
# Finish
```

Нестандартные директивы

- Директива **Crawl-delay** (Yahoo и MSN) – время в секундах между запросами робота.
- Директива **Allow** (Yahoo и Google) – указывает адреса, которые можно индексировать
- Символы подстановки * - любые символы и **$** - конец строки (Yahoo, Google, Rambler)
- Директива **Host** (Yandex) – директива указывает на главное зеркало сайта

Пример файла robots.txt

```
User-agent: msnbot-media
User-agent: Googlebot-Image
User-agent: Yahoo-MMCrawler
Disallow: /

User-agent: Yandex
Disallow: /Messages.asp?sort=
Host: forum.liga.net

User-agent: Googlebot
User-agent: StackRambler
Disallow: /*ts=
Disallow: /*=$

User-agent: Slurp
User-agent: msnbot
Disallow: /Messages.asp?sort=
Crawl-delay: 10

User-agent: *
Disallow: /Messages.asp?sort=
Disallow: /poll/
Disallow: /ic
```

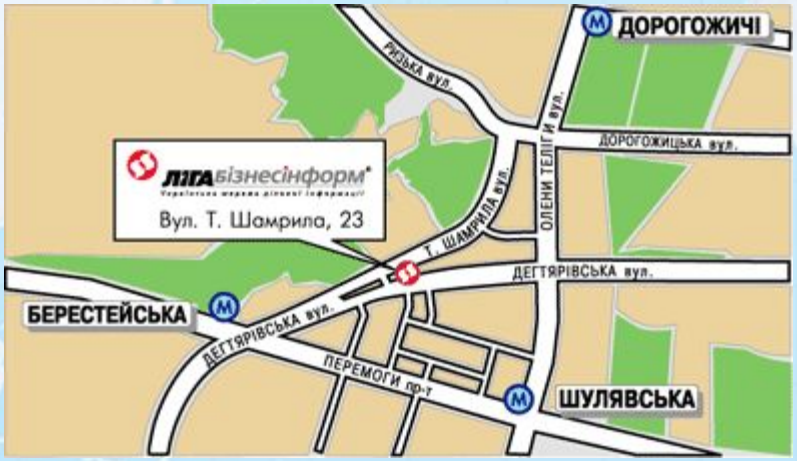
Резюме

- Создавать robots.txt для каждого домен и поддомена сайта
- Создавать запись для всех остальных роботов (User-agent: *)
- Использовать нестандартные директивы только в секциях для тех роботов, которые их поддерживают

ЛІГАбізнесінформ
Українська мережа ділової інформації

www.liga.net

Информационно-аналитический центр «ЛІГА»



04112, г. Киев,
ул. Т. Шамрыло, 23

Тел./факс: +380 (44) 538-01-01
(многоканальный)

E-mail: marketing@liga.net
Web: www.liga.net

Поисковая оптимизация и продвижение сайтов в Интернете

Владимир Чернышов