Курс SEO-практик

Управление индексацией сайта. Дубли и служебные страницы

Модуль 7

IT-Academy

bit.ly/2JKmiS0

Задание для самостоятельного выполнения

Самостоятельно разобраться с программой Xenu или другой на выбор.

http://stalnik.by/

- ✔ проверить на наличие «битых» ссылок и редиректов с помощью выбранной программой
- ✔ разобраться в возможной причине
- ✔ постараться дать рекомендации по исправлению

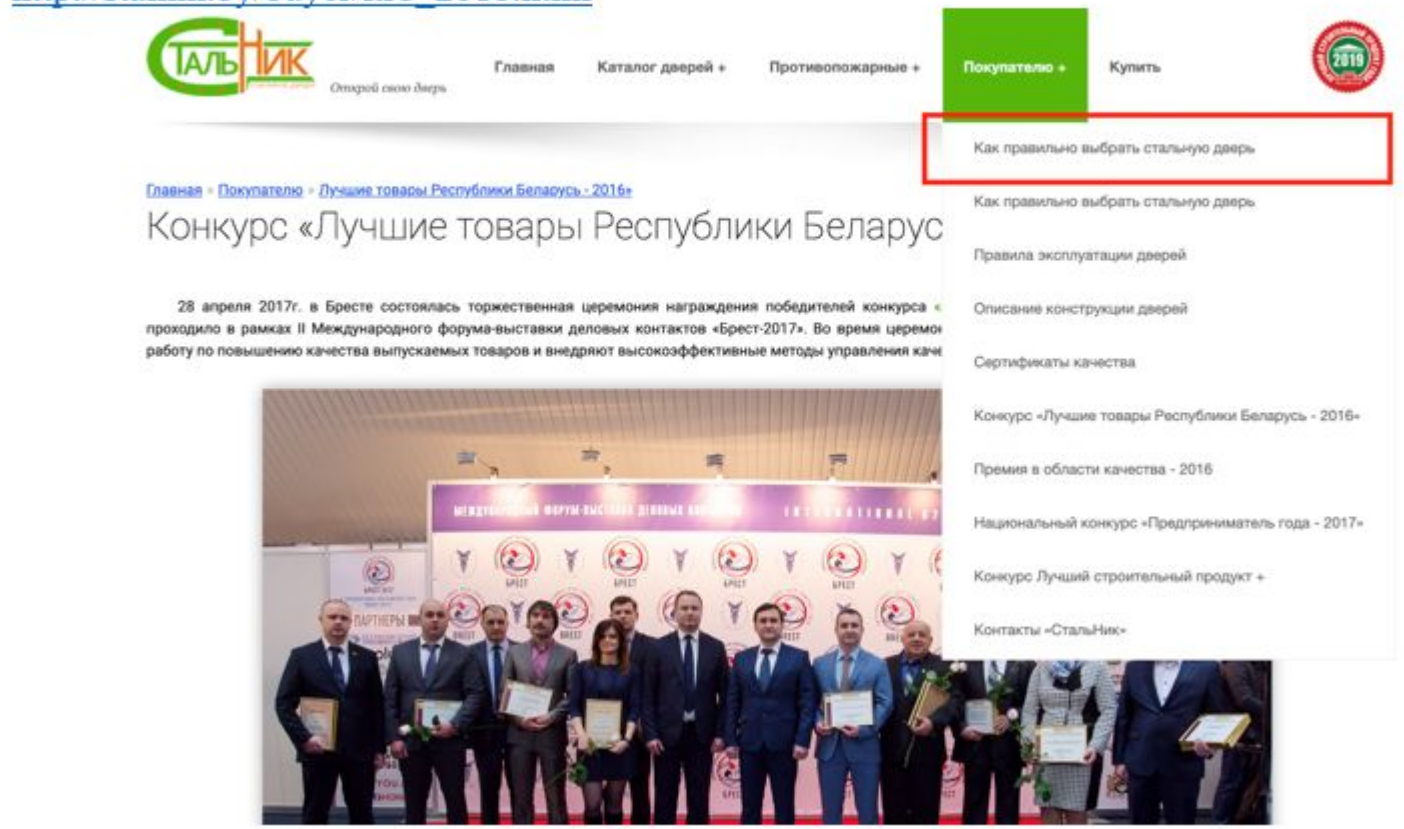
Разбор

1. http://stalnik.by/buyer/choice.html
2. http://stalnik.by/buyer/contacts.html
3. http://stalnik.by/buyer/fire.html
4. http://stalnik.by/buyer/fire2.html
5. http://stalnik.by/buyer/pvokbo_2018.html
6. http://stalnik.by/doors/buyer/choice.html
7. http://stalnik.by/doors/stock/buyer/choice.html
8. http://stalnik.by/style-content/searchicon.png
9. http://stalnik.by/style-content/slider/assets/transparent.png

Разбор

1. http://stalnik.by/buyer/buyer/choice.html - 404

Этой страницы не существует, ее нужно удалить или сделать. На нее веду ссылки со следующих страниц:
http://stalnik.by/buyer/ltrb_2016.html



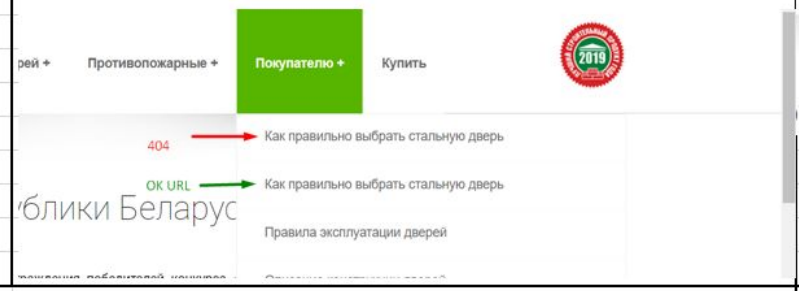
Эту ссылку надо удалить или перелинковать на существующую страницу

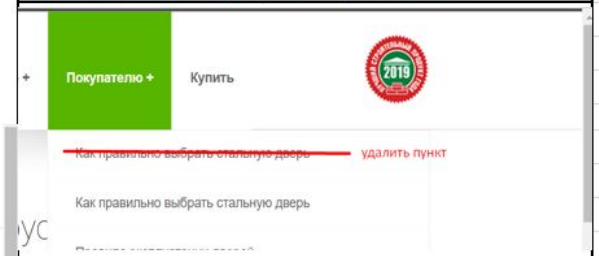
Разбор

http://stalnik.by/buyer/buyer/choice.html	404 Not Found	В меню неверно указана ссылка на страницу Редикт на страницу http://stalnik.by/buyer/choice.html
http://stalnik.by/buyer/buyer/contacts.html	404 Not Found	В меню неверно указана ссылка на страницу Редикт на страницу http://stalnik.by/buyer/contacts.html
http://stalnik.by/buyer/fire.html	404 Not Found	Неверно указана ссылка на страницу Редикт на страницу http://stalnik.by/fire.html
http://stalnik.by/buyer/fire2.html	404 Not Found	Неверно указана ссылка на страницу Редикт на страницу http://stalnik.by/fire2.html
http://stalnik.by/doors/buyer/choice.html	404 Not Found	Неверно указана ссылка на страницу Редикт на страницу http://stalnik.by/buyer/choice.html
http://stalnik.by/buyer/pvokbo_2018.html	404 Not Found	Неверно указана ссылка на страницу Редикт на страницу http://stalnik.by/buyer/pvokbo_2016.html

Разбор

	A	B	C	D
1			Рекомендация по исправлению	
2	http://stalnik.by/buyer/buyer/choice.html	на сайте в некоторых разделах, присутствует дубль ссылки на одну и туже статью "Как выбрать входную дверь". Развернутый пукт меню "Покупателю +" на таких страницах включает в свою структуру 2 одинаковых загаловка (рис.1). Один из них имеет верную ссылку для перехода, а другой - "битую" ссылку и ведет на страницу "404 Not Found"		
3	error code: 404 (not found), linked from page(s):		http://stalnik.by/buyer/ltrb_2016.html	удалить лишний пункт меню с битой ссылкой
4	http://stalnik.by/buyer/ltrb_2016.html		http://stalnik.by/buyer/nkpg_2017.html	удалить лишний пункт меню с битой ссылкой
5	http://stalnik.by/buyer/nkpg_2017.html			
6				
7				
8				
9				
10				
11				
12				
13				
14				
15				





Разбор

На мой взгляд, указанные на листе "404 ошибки" очень похожи на ошибки разработчиков, т.к. они почти все достаточно типовые.

Однако, как мне кажется, тут может быть вопрос с их возникновением, т.к. такие ошибки могли появится вследствие изменения структуры сайта (к примеру часто встретилась ошибка в ссылках формата /buyer/buyer)

Т.е. теоретически, они могли появится из-за того, что был раздел+подраздел, а затем подраздел был удален.

Я дал рекомендации исходя из первого предположения (ошибок разработчиков). Следовательно исходил из того, что таких же внешних ссылок, ведущих на 404, быть не должно.

Как я понимаю, в любом случае этот момент нужно уточнять с разработчиками, т.к. если ошибки связаны с изменением структуры, то тогда необходимо в тех пунктах, где указано удаление и исправление ссылок, делать 301 редирект.

Разбор

	A	B
1	http://www.stalnik.by/	редирект с http://www.stalnik.by/ на главное зеркало http://stalnik.by/
2	redirected to: http://stalnik.by/	
3	status code: 301 (object permanently moved)	

Также с http://www.stalnik.by/ - стоит 301 редирект на без WWW

Разбор

2. 301 и 302 редиректы:

http://www.stalnik.by/
redirected to: http://stalnik.by/

status code: 301 (object permanently moved)
linked from page(s):
http://stalnik.by/
http://stalnik.by/doors/stock.html
http://stalnik.by/buyer/rules.html
http://stalnik.by/buyer/construction.html
http://stalnik.by/buyer/ltrb_2016.html
http://stalnik.by/doors/flat.html
http://stalnik.by/doors/house_two.html
http://stalnik.by/buyer/certificate.html
http://stalnik.by/doors/house.html
http://stalnik.by/buyer/choice.html
http://stalnik.by/buyer/pvokbo_2016.html
http://stalnik.by/buyer/lspg_2017.html
http://stalnik.by/buyer/nkpg_2017.html
http://stalnik.by/buyer/lspg_2015.html
http://stalnik.by/fire2.html

Разбор

Ошибка 301 редиректа решается путем подключения к сайту по протоколу FTP, затем в корневой категории сайта найти файл .htaccess. И добавить в файл следующий код:

RewriteRule ^aksessuary/powerbank$ /gadzhety-aksessuary/powerbank [R=301,L]

Ссылки с ошибкой с сервера 404	Страницы, на которых указан неправильный путь	Правильный путь, где лежит страница (нужно прописать его
http://stalnik.by/buyer/buyer/choice.html	http://stalnik.by/buyer/ltrb_2016.html http://stalnik.by/buyer/nkpg_2017.html	http://stalnik.by/buyer/choice.html
http://stalnik.by/buyer/buyer/contacts.html	http://stalnik.by/buyer/choice.html http://stalnik.by/buyer/rules.html http://stalnik.by/buyer/construction.html http://stalnik.by/buyer/certificate .html http://stalnik.by/buyer/ltrb_2016.html http://stalnik.by/buyer/pvokbo_2016.html http://stalnik.by/buyer/nkpg_2017.html http://stalnik.by/buyer /lspg_2015.html http://stalnik.by/buyer/lspg_2016.html http://stalnik.by/buyer/lspg_2017.html http://stalnik.by/buyer/lspg_2018.html http://stalnik.by /buyer/contacts.html	http://stalnik.by/buyer/contacts.html
http://stalnik.by/buyer/fire.html	http://stalnik.by/buyer/ltrb_2016.html http://stalnik.by/buyer/nkpg_2017.html	http://stalnik.by/fire.html
http://stalnik.by/buyer/fire2.html	http://stalnik.by/buyer/ltrb_2016.html http://stalnik.by/buyer/nkpg_2017.html	http://stalnik.by/fire2.html
http://stalnik.by/doors/buyer/choice.html	http://stalnik.by/doors/stock.html http://stalnik.by/doors/flat.html http://stalnik.by/doors/house.html http://stalnik.by/doors/house -by-name_two.html http://stalnik.by/doors/house-by-name.html	http://stalnik.by/buyer/choice.html
http://stalnik.by/doors/stock/buyer/choice.html	http://stalnik.by/doors/stock/optima.html http://stalnik.by/doors/stock/ritm.html http://stalnik.by/doors/stock/vector.html http: // stalnik .by / doors / stock / logika.html http://stalnik.by/doors/stock/universal.html	http://stalnik.by/buyer/choice.html

Курс SEO-практик

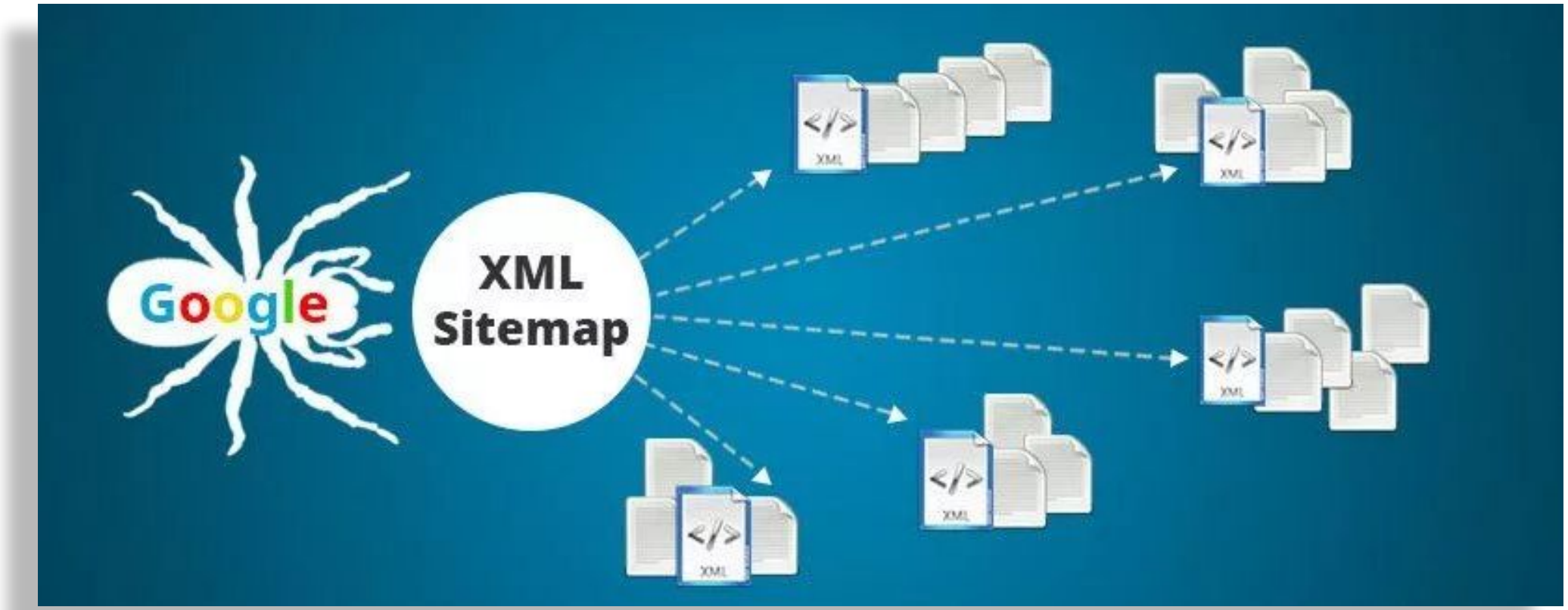
Управление индексацией сайта. Дубли и служебные страницы

Модуль 7

IT-Academy

Sitemap.xml для чего необходим и как создать

Sitemap.xml



Sitemap.xml для чего необходим и как создать

Sitemap.xml

– карта сайта в формате XML, которая содержит ссылки на все разделы и страницы сайта подлежащие индексации.

Альтернативное название: XML карта сайта

Файл Sitemap.xml позволяет сообщить поисковым системам о том, как организован контент на вашем сайте. Поисковые роботы просматривают этот файл, чтобы более точно индексировать ваши страницы.

Sitemap.xml для чего необходим и как создать

Нужен ли файл Sitemap.xml?

Если страницы файла корректно связаны друг с другом, поисковые роботы могут обнаружить большую часть материалов. Тем не менее, с помощью файла Sitemap можно оптимизировать сканирование сайта, особенно в следующих случаях:

✔**Размер сайта очень велик.**

✔**Сайт содержит большой архив страниц, которые не связаны друг с другом.** Чтобы они были успешно просканированы, их можно перечислить в файле Sitemap.

✔**Сайт создан недавно, и на него указывает мало ссылок.** Робот Googlebot и другие поисковые роботы сканируют Интернет, переходя по ссылкам с одной страницы на другую. Если на ваш сайт указывает мало ссылок, его будет сложного найти.

Sitemap.xml для чего необходим и как создать

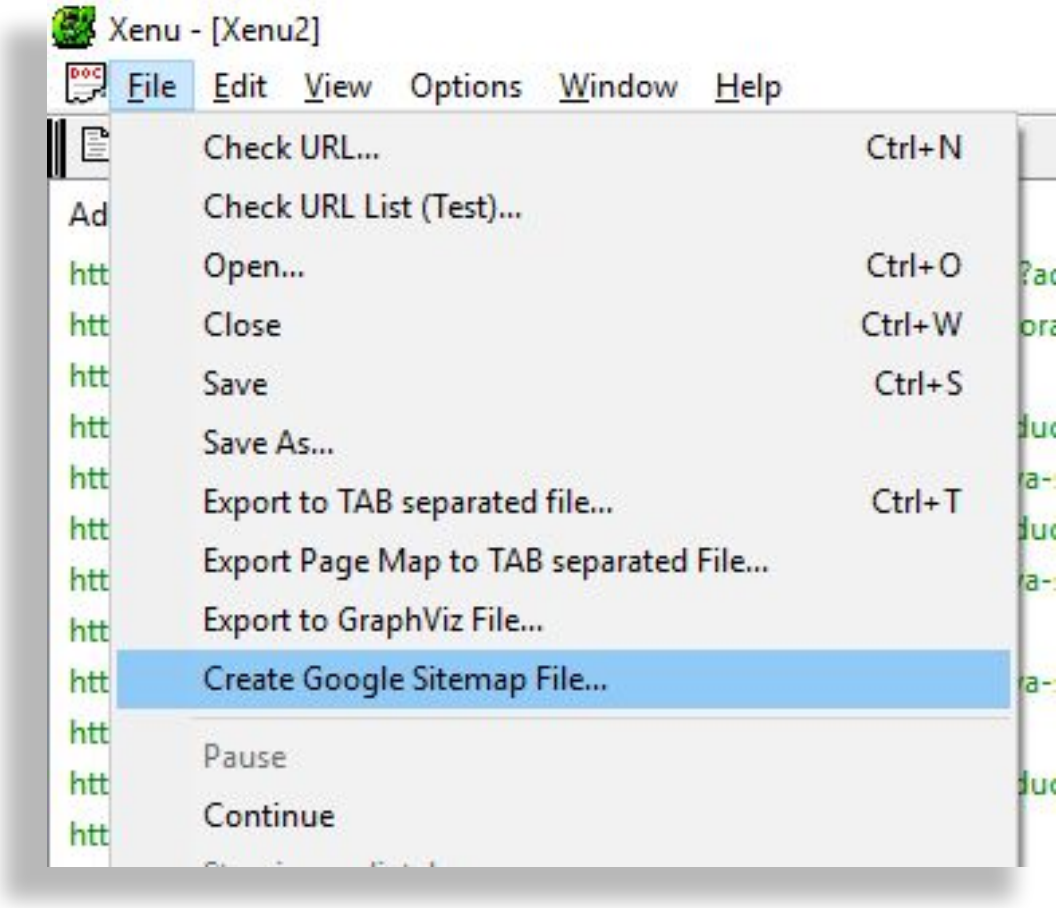
Как создать Sitemap.xml

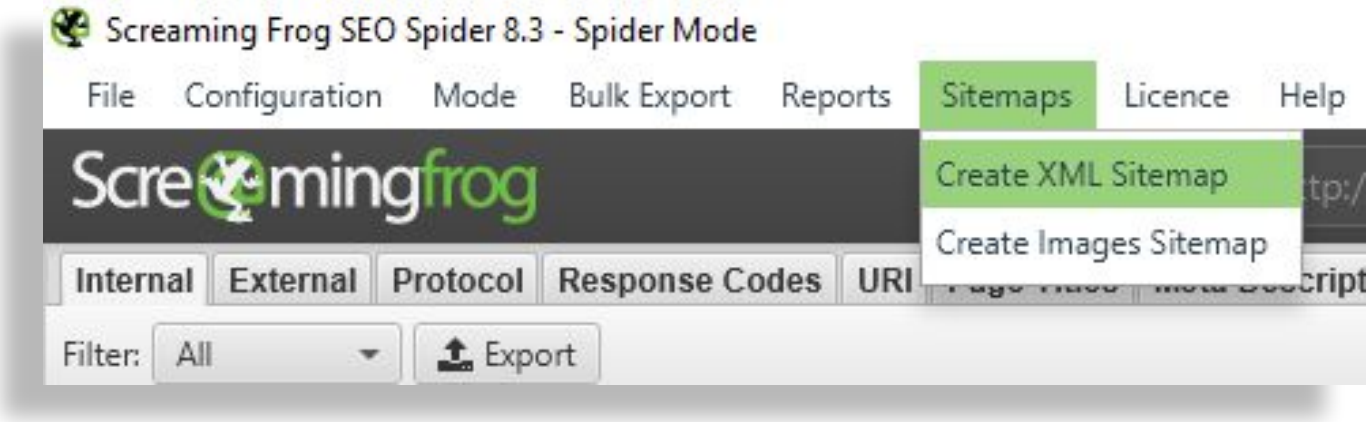
Генерация средствами CMS

Генерация сторонними сервисами\программами

- ✔ http://www.mysitemapgenerator.com/ (до 500 страниц бесплатно)
- ✔ Xenu
- ✔ Screaming Frog SEO Spider

Sitemap.xml для чего необходим и как создать





Sitemap.xml для чего необходим и как создать

Синтаксис для Sitemap.xml

Яндекс и Google поддерживают стандартный протокол Sitemap

https://www.sitemaps.org/ru/protocol.html

Sitemap.xml для чего необходим и как создать

Обязательные атрибуты:

<urlset> - определяет стандарт протокола и инкапсулирует этот файл.

<url> - Родительский тег для каждой записи URL-адреса. Остальные теги являются дочерними для этого тега.

<loc> - URL-адрес страницы. Этот URL-адрес должен начинаться с префикса (например, HTTP) и заканчиваться косой чертой, если Ваш веб-сервер требует этого. Длина этого значения не должна превышать 2048 символов.

Sitemap.xml для чего необходим и как создать

Необязательные атрибуты:

<lastmod> - Дата последнего изменения файла.

<changefreq> - Вероятная частота изменения этой страницы. Это значение предоставляет общую информацию для поисковых систем и может не соответствовать точно частоте сканирования этой страницы.

<priority> - Приоритетность URL относительно других URL на Вашем сайте. Допустимый диапазон значений — от 0,0 до 1,0.

Sitemap.xml для чего необходим и как создать

Пример sitemap.xml

https://www.termebel.by/sitemap.xml (1)

Sitemap.xml наиболее частые ошибки

Основные требования Google и Яндекса:

✔Используйте кодировку UTF-8.

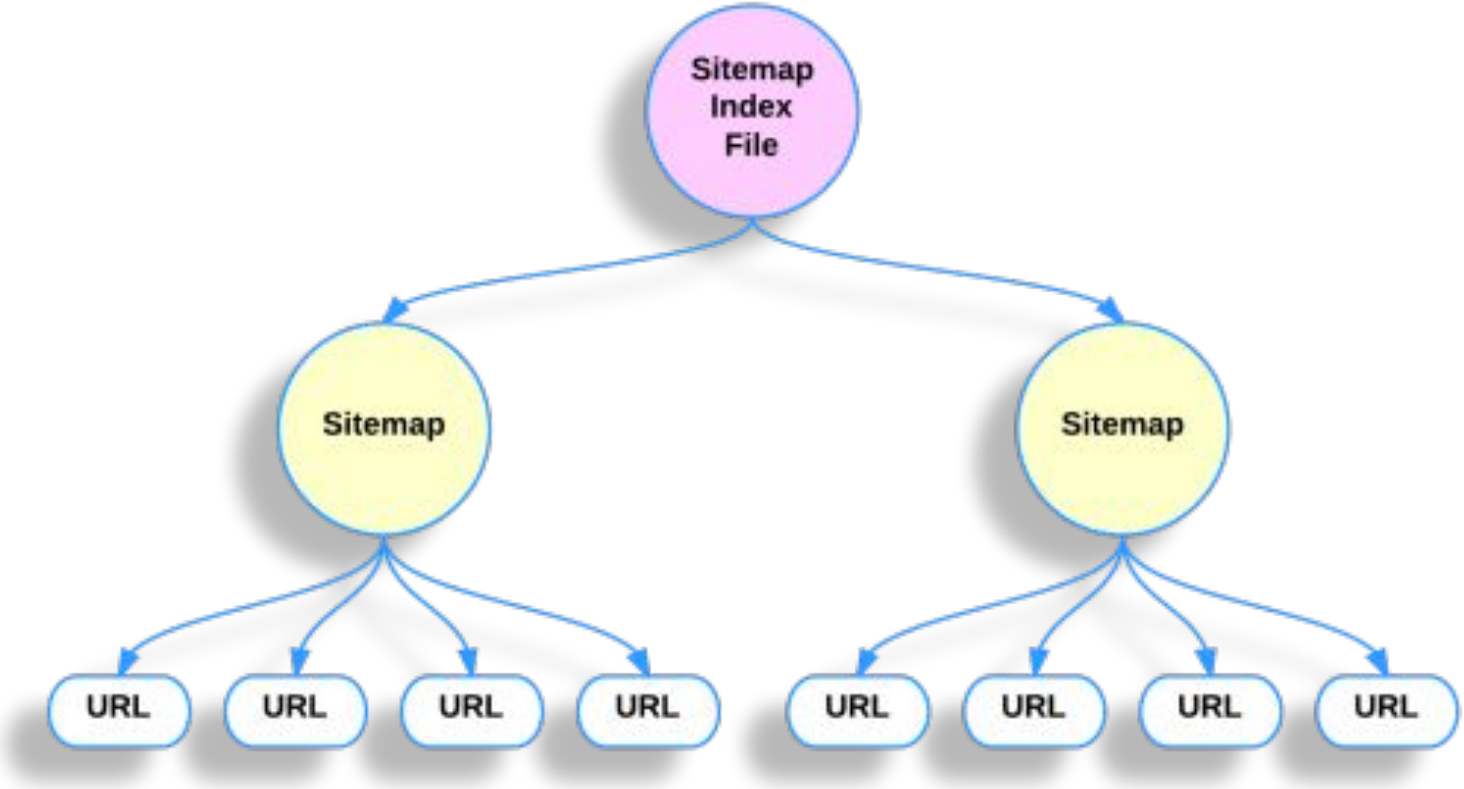
✔Максимальное количество ссылок — 50 000. Вы можете разделить Sitemap на несколько отдельных файлов и указать их в файле индекса Sitemap.

✔Указывайте ссылки на страницы только того домена, на котором будет расположен файл.

✔Разместите файл на том же домене, что и сайт, для которого он составлен.

✔При обращении к файлу сервер должен возвращать HTTP-код 200.

Sitemap.xml наиболее частые ошибки



Пример: https://wilmax24.by/sitemap.xml (2)

Sitemap.xml наиболее частые ошибки

Отличия:

Рекомендации Яндекса к файлу:

✔ Поддерживает кириллические URL.

Рекомендации Google:

✔ Поддерживает только цифры и латинские буквы.

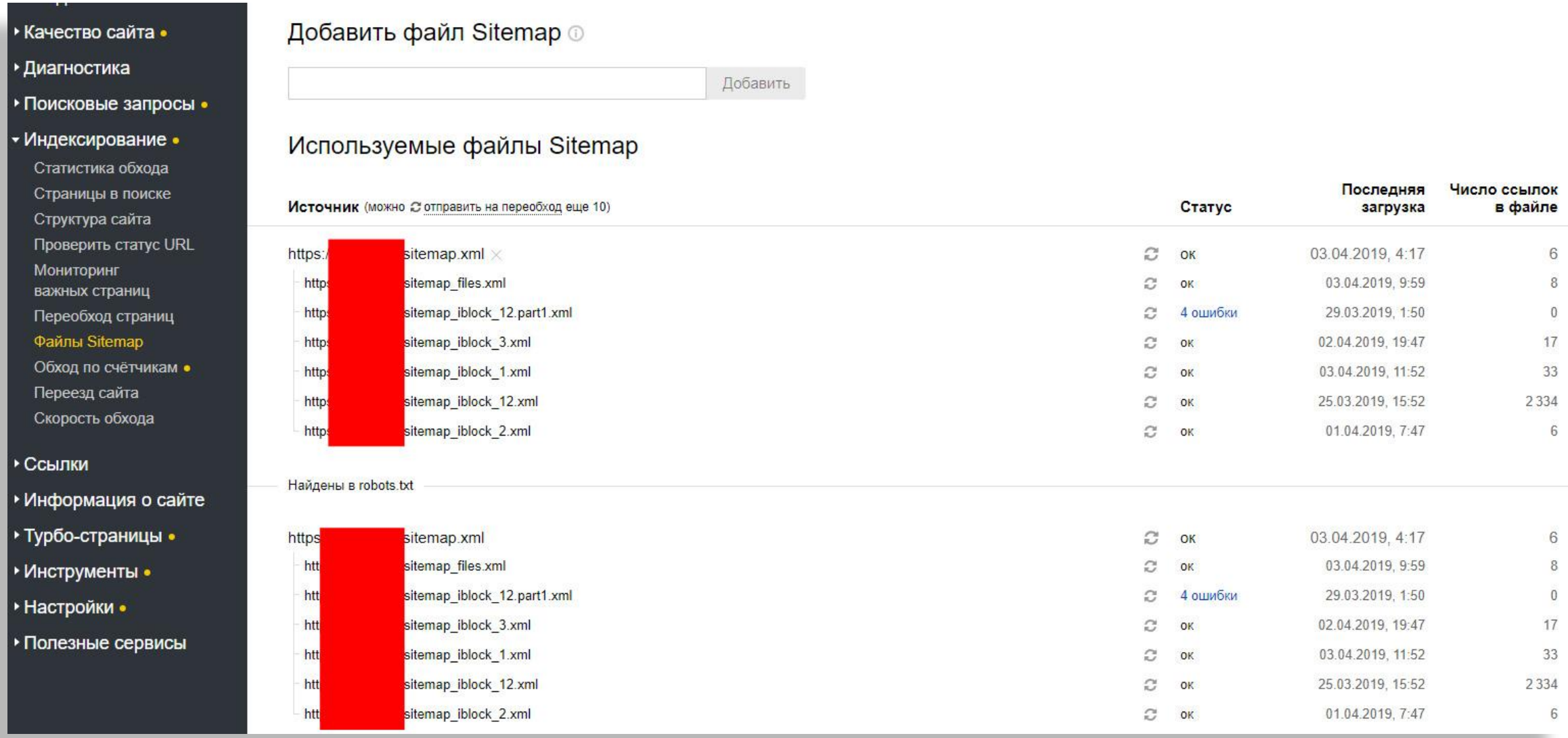
Sitemap.xml наиболее частые ошибки

Как сообщить поисковым системам о Sitemap.xml:

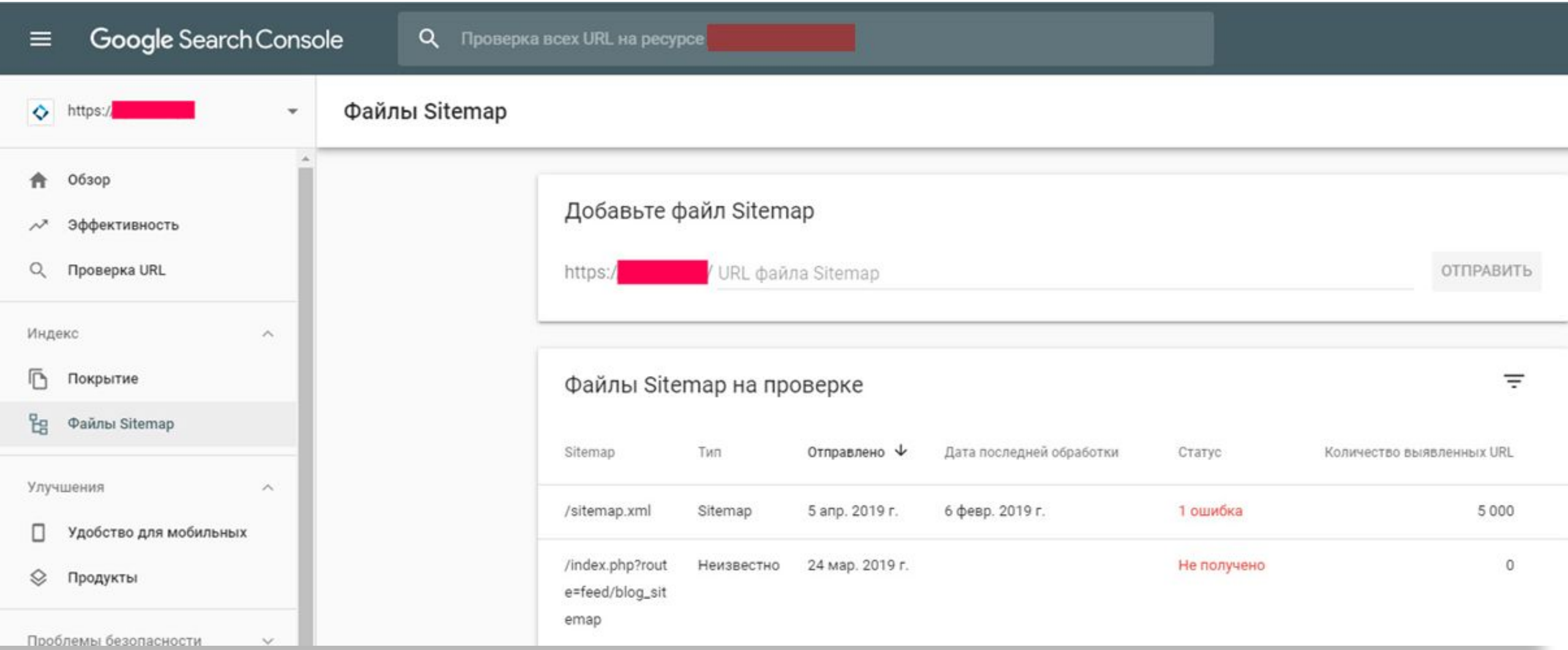
- ✔ Укажите ссылку на файл в robots.txt
- ✔ Добавить Sitemap.xml через Яндекс.Вебмастер и Google Search Console

Важно! Можно выбрать 1 из способов.

Sitemap.xml наиболее частые ошибки



Sitemap.xml наиболее частые ошибки



Sitemap.xml наиболее частые ошибки

Наиболее частые ошибки:

- ✔ Нет регулярной актуализации Sitemap.xml;
- ✔ Содержит ссылки на 404 и 301 страницы;
- ✔ Содержит ссылки на страницы с ответом сервера 200, которые не подлежат индексации;
- ✔ Google и Яндекс не знают о существовании sitemap.xml.

Sitemap.xml наиболее частые ошибки

Частые заблуждения:

- ✔ Включение URL-адреса в файл Sitemap.xml гарантирует, что он будет проиндексирован;
- ✔ Если удалить URL из Sitemap.xml, он будет удалён из индекса;
- ✔ Sitemap.xml трудно создавать и поддерживать.
- ✔ Sitemap.xml должен быть только по URL domen.by/sitemap.xml

Sitemap.xml наиболее частые ошибки

Google и Яндекс поддерживают не только формат XML для Sitemap:

https://support.google.com/webmasters/answer/183668?hl=ru

https://yandex.ru/support/webmaster/controlling-robot/sitemap.html#sitemap__yandex-supported-formats

Sitemap.xml наиболее частые ошибки

Проверить корректность Sitemap.xml (синтаксис):

Если нет доступа к панелям вебмастеров (например, сайт еще там не зарегистрирован, либо нет к ним доступа), то можно использовать:

https://webmaster.yandex.ru/tools/sitemap/ (3) (не требует регистрации в Яндекс.Вебмастере)

Robots.txt директивы и их использование

robots.txt

Robots.txt директивы и их использование

Robots.txt

- текстовый файл, который содержит параметры индексирования сайта для роботов поисковых систем.

Файл должен располагаться в корневом каталоге в виде обычного текстового документа и быть доступен по адресу: https://site.by/robots.txt.

Robots.txt директивы и их использование

Зачем нужен файл robots.txt

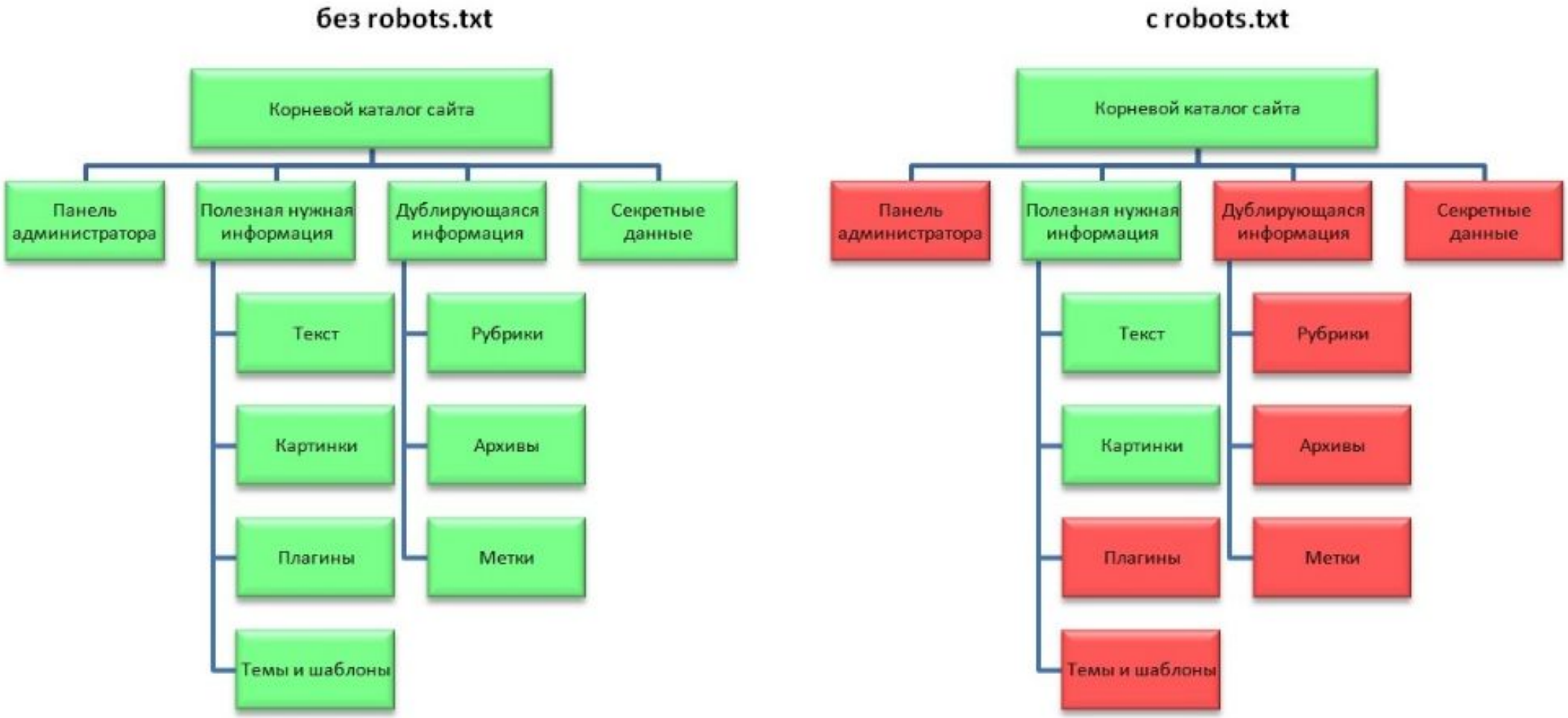
Например, мы не хотим, чтобы роботы поисковых систем посещали:

- ✔ страницы с личной информацией пользователей на сайте;
- ✔ страницы с разнообразными формами отправки информации;
- ✔ страницы с результатами поиска.

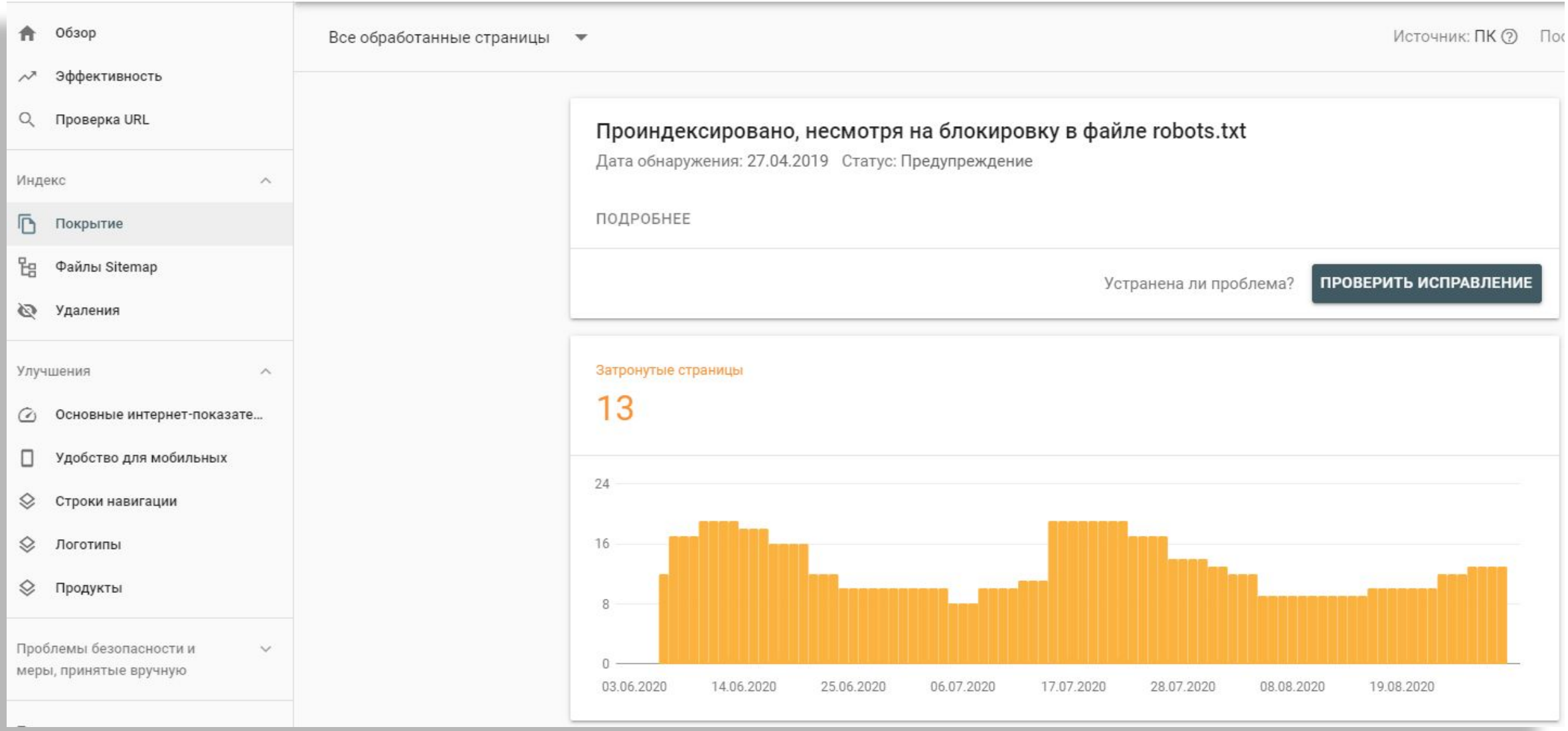
Важно понимать, что закрытие страницы не является 100% гарантией того, что робот ее не проиндексирует!

Robots.txt директивы и их использование

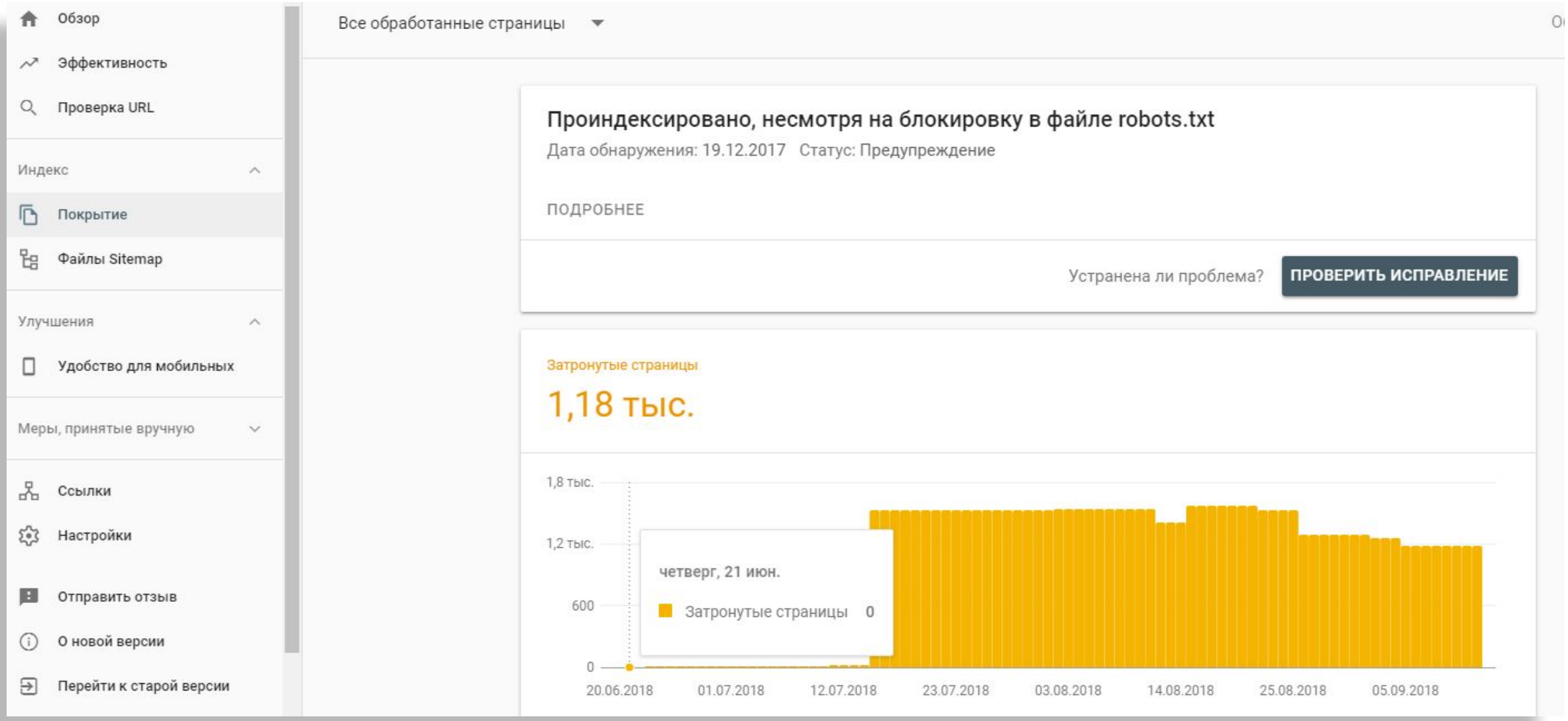
Пример того, что индексируется поисковиками:



Robots.txt директивы и их использование



Robots.txt директивы и их использование



Robots.txt директивы и их использование

Директива robots.txt

– это инструкция, которая обрабатывается роботами поисковых систем.

Какие директивы бывают:

User-agent

Disallow и Allow

Sitemap

Host (уже неактуальна, но часто встречается до сих пор)

Crawl-delay

Clean-param

Robots.txt директивы и их использование

User-agent

- правило о том, каким роботам необходимо просмотреть инструкции, описанные в файле robots.txt.

User-agent: *

User-agent: Googlebot

User-agent: Yandex

Robots.txt директивы и их использование

Disallow: - чтобы запретить доступ робота к сайту, некоторым его разделам или страницам

User-agent: *

Disallow: / # блокирует доступ ко всему сайту

User-agent: *

Disallow: /bin # блокирует доступ к страницам, #начинающимся с '/bin'

Robots.txt директивы и их использование

Allow: - чтобы разрешить доступ робота к сайту, некоторым его разделам или страницам

User-agent: Yandex

Allow: /cgi-bin

Disallow: /

запрещает скачивать все, кроме страниц, начинающихся с '/cgi-bin'

Robots.txt директивы и их использование

Директивы Allow и Disallow из соответствующего User-agent блока сортируются по длине префикса URL (от меньшего к большему) и применяются последовательно. Если для данной страницы сайта подходит несколько директив, то робот выбирает последнюю в порядке появления в сортированном списке. Таким образом, порядок следования директив в файле robots.txt не влияет на использование их роботом.

Robots.txt директивы и их использование

Исходный robots.txt:
User-agent: Yandex
Allow: /
Allow: /catalog/auto
Disallow: /catalog
Сортированный robots.txt:
User-agent: Yandex
Allow: /
Disallow: /catalog
Allow: /catalog/auto
запрещает скачивать страницы, начинающиеся с '/catalog', но разрешает #скачивать страницы, начинающиеся с '/catalog/auto‘ и остальные.

Robots.txt директивы и их использование

Директивы Allow и Disallow без параметров

```
User-agent: *
Disallow:          # то же, что и Allow: /

User-agent: *
Allow:             # не учитывается роботом
```

Robots.txt директивы и их использование

При указании путей директив Allow и Disallow можно использовать спецсимволы * и $, задавая, таким образом, определенные регулярные выражения. Спецсимвол * означает любую (в том числе и отсутствие) последовательность символов.

```
User-agent: *
Disallow: /cgi-bin/*.aspx      # запрещает '/cgi-bin/example.aspx'
                               # и '/cgi-bin/private/test.aspx'
Disallow: /*private            # запрещает не только '/private',
                               # но и '/cgi-bin/private'
```

Robots.txt директивы и их использование

User-agent: *
Disallow: /catalog/*.html

site.by/catalog/tv/
site.by/catalog/tv/Samsung.html

Disallow: /*tv
site.by/catalog/Tv/
site.by/catalog/tv/
site.by/catalog/smart-tv/Samsung.html

Robots.txt директивы и их использование

По умолчанию к концу каждого правила, описанного в файле robots.txt, приписывается спецсимвол *. Пример:

User-agent: *

Disallow: /catalog* #блокирует доступ к страницам, #начинающимся с '/catalog'

Disallow: /catalog #то же самое

Robots.txt директивы и их использование

Чтобы отменить * на конце правила, можно использовать спецсимвол $, например:

User-agent: Yandex

Disallow: /tv/$

site.by/tv/

site.by/tv/Samsung.html

Robots.txt директивы и их использование

Использование кириллицы запрещено

Для указания имен доменов используйте Punycode https://ru.wikipedia.org/wiki/Punycode

#Неверно:

User-agent: Yandex

Disallow: /корзина

#Верно:

User-agent: Yandex

Disallow: /%D0%BA%D0%BE%D1%80%D0%B7%D0%B8%D0%BD%D0%B0

Robots.txt директивы и их использование

Директива Sitemap

User-agent: *

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

Важно указывать полный путь с указанием протокола!

Robots.txt директивы и их использование

Директива Host: ранее использовалась для указания главного зеркала сайта, учитывалась только Яндексом. Теперь и он ее не учитывает.

User-Agent: *

Host: https://site.by

Robots.txt директивы и их использование

Директива Crawl-delay - Если сервер сильно нагружен и не успевает отрабатывать запросы на загрузку. Она позволяет задать поисковому роботу минимальный период времени (в секундах) между окончанием загрузки одной страницы и началом загрузки следующей.

User-agent: Yandex

Crawl-delay: 2.0 # задает таймаут в 2 секунды

Google не учитывает!

Robots.txt директивы и их использование

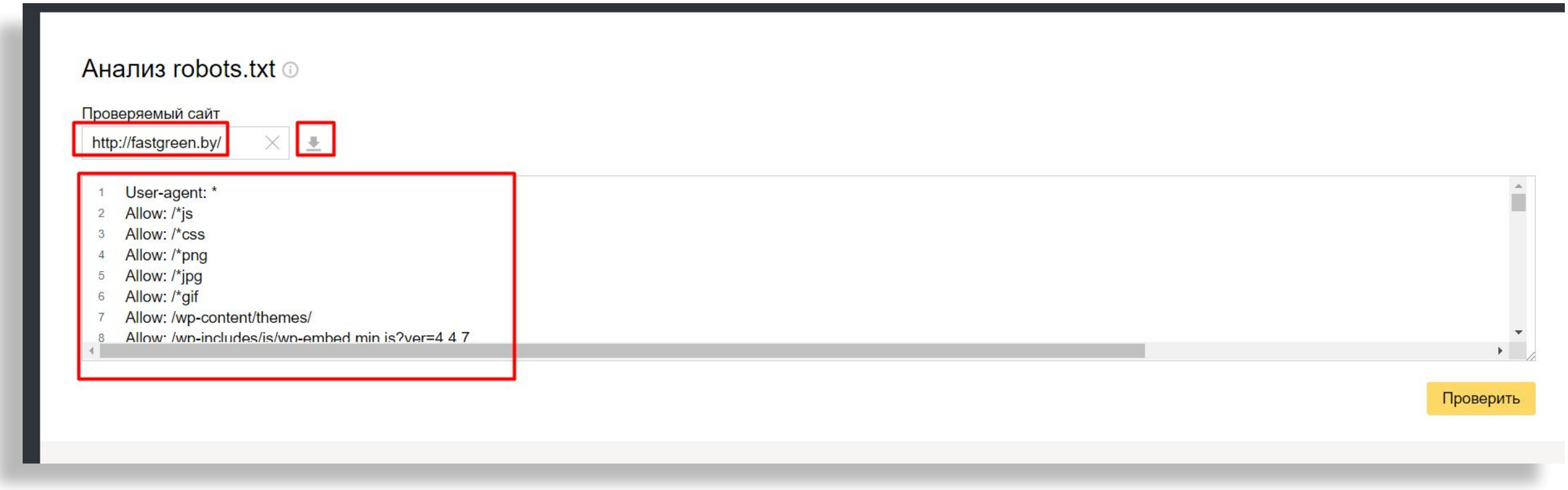
Директива Clean-param

- Если адреса страниц сайта содержат динамические параметры, которые не влияют на их содержимое (например: идентификаторы сессий, пользователей, рефереров и т. п.), вы можете описать их с помощью директивы Clean-param.

Robots.txt директивы и их использование

https://webmaster.yandex.ru/tools/robotstxt/ (4)- проверка robots.txt

Robots.txt директивы и их использование



Robots.txt директивы и их использование

Результаты анализа robots.txt

0 ошибок | используемые строки

Строки	Используемые секции
29	User-agent: Yandex
	Allow: /*js
	...
55	Disallow: /page/
57	Host: fastgreen.by
58	Sitemap: http://fastgreen.by/sitemap.xml

Robots.txt директивы и их использование

Разрешены ли URL?

Список URL

```
1 http://fastgreen.by/service/
```

Проверить

URL	Результат
http://fastgreen.by/service/	✓

Robots.txt директивы и их использование

Практическое задание

Robots.txt директивы и их использование

Для сайта https://linenmill.by (5) доработать текущий robots.txt с учетом необходимости закрытия следующих страниц от индексации ПС Яндекс.

https://linenmill.by/kontraktnyj-zakaz/ (a)

https://linenmill.by/author/vova/ (b)

https://linenmill.by/author/zenya/ (c)

Проверить корректность в https://webmaster.yandex.ru/tools/robotstxt/

Robots.txt директивы и их использование

Добавили в блок «User-agent: Yandex» следующие директивы:

Disallow: /kontraktnyj-zakaz/$

Disallow: /author/vova/$

Disallow: /author/zenya/$

Получили:

URL	Результат
https://linenmill.by/kontraktnyj-zakaz/	/kontraktnyj-zakaz/$
https://linenmill.by/author/vova/	/author/vova/$
https://linenmill.by/author/zenya/	/author/zenya/$

Базовые условия индексации документа, проверка индексации

Базовые условия индексации документа, проверка индексации

- ✔ Страница должна отдавать код ответа сервера 200 ОК;
- ✔ Страница не запрещена для индексирования в файле robots.txt;
- ✔ Страница не является дублем другой страницы в рамках сайта;
- ✔ Страница содержит полезный контент, и может быть полезна пользователям;

Базовые условия индексации документа, проверка индексации

Проверка индексации:

- ✔ Информация в панелях вебсмастеров Яндекса и Google
- ✔ Запросы с использованием операторов

url:site.by/catalog/page1.html - Яндекс для страницы

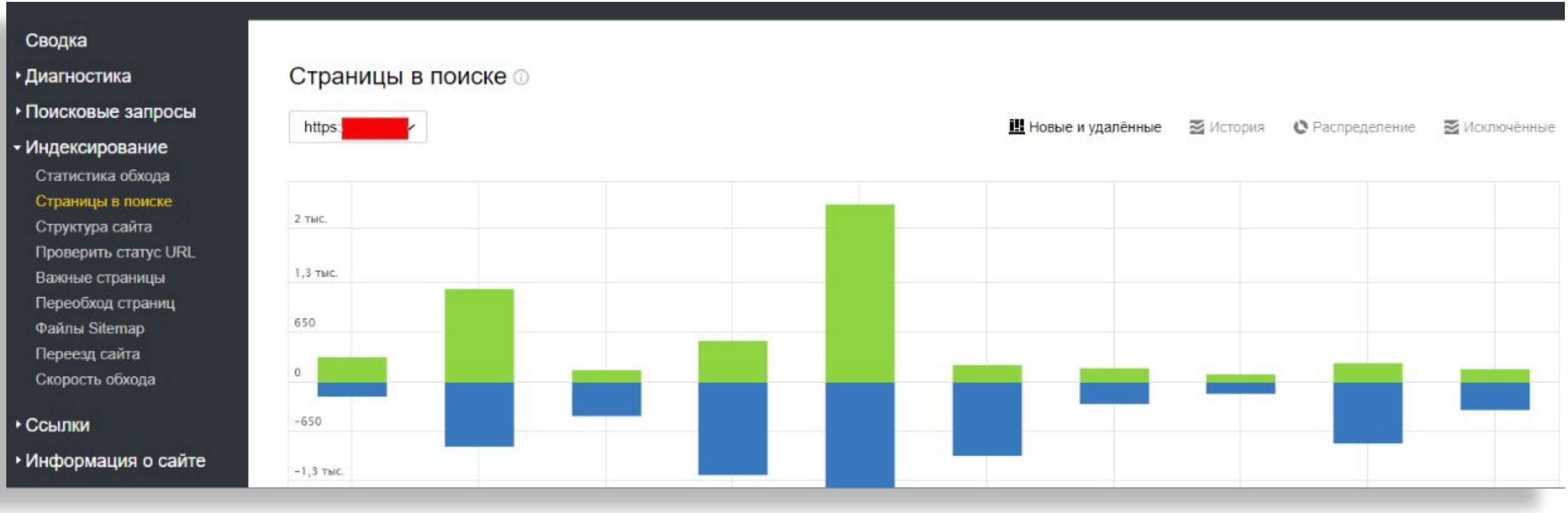
url:site.by/* - Яндекс для сайта

~~info:https://site.by/catalog/page1.html – Google для страницы~~

site:site.by – Google для сайта

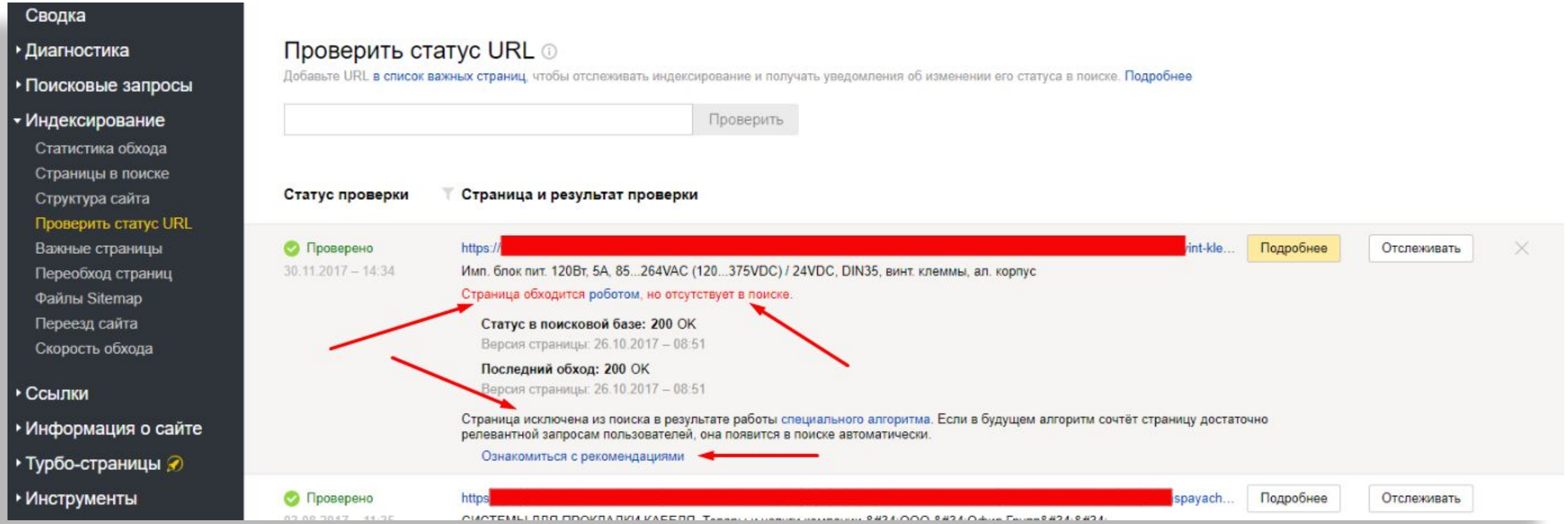
- ✔ Плагины для браузера, например, RDS bar

Базовые условия индексации документа, проверка индексации

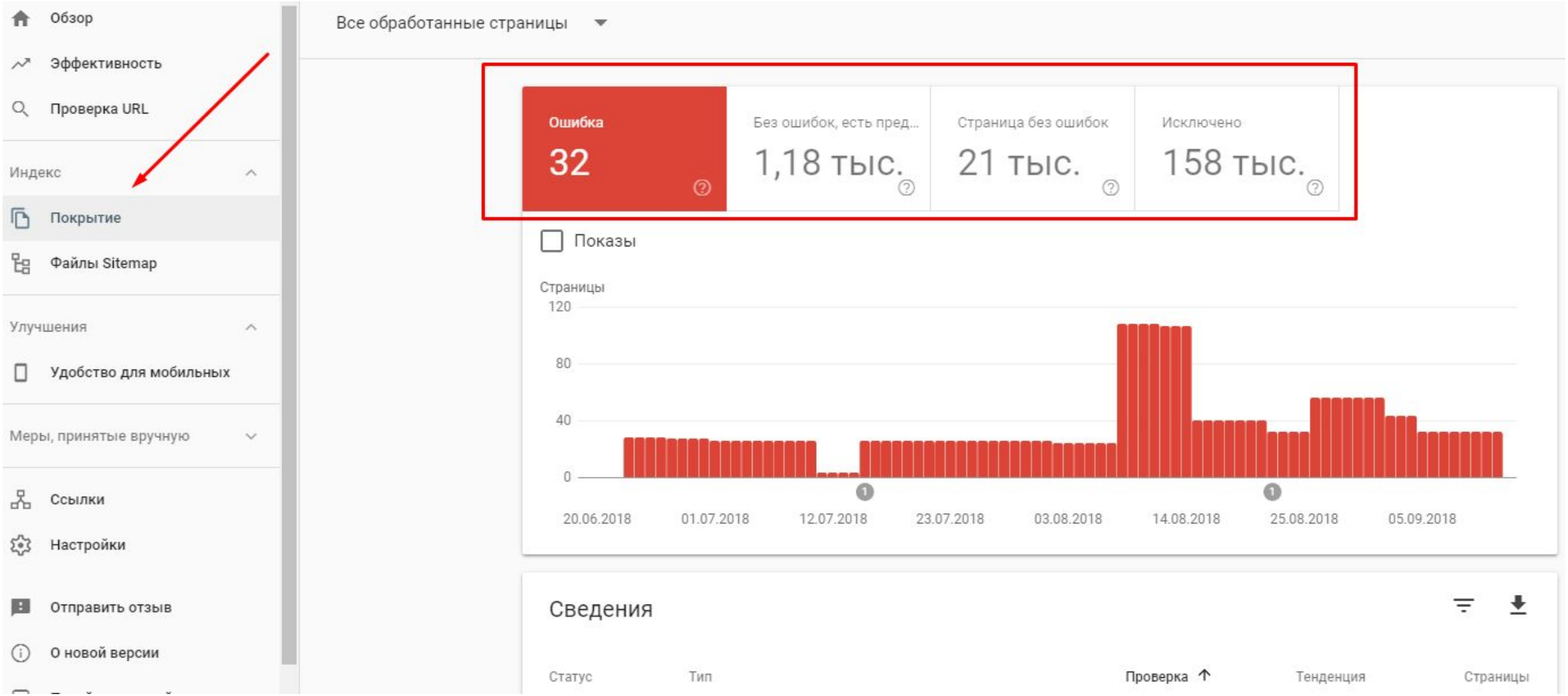


Базовые условия индексации документа, проверка индексации

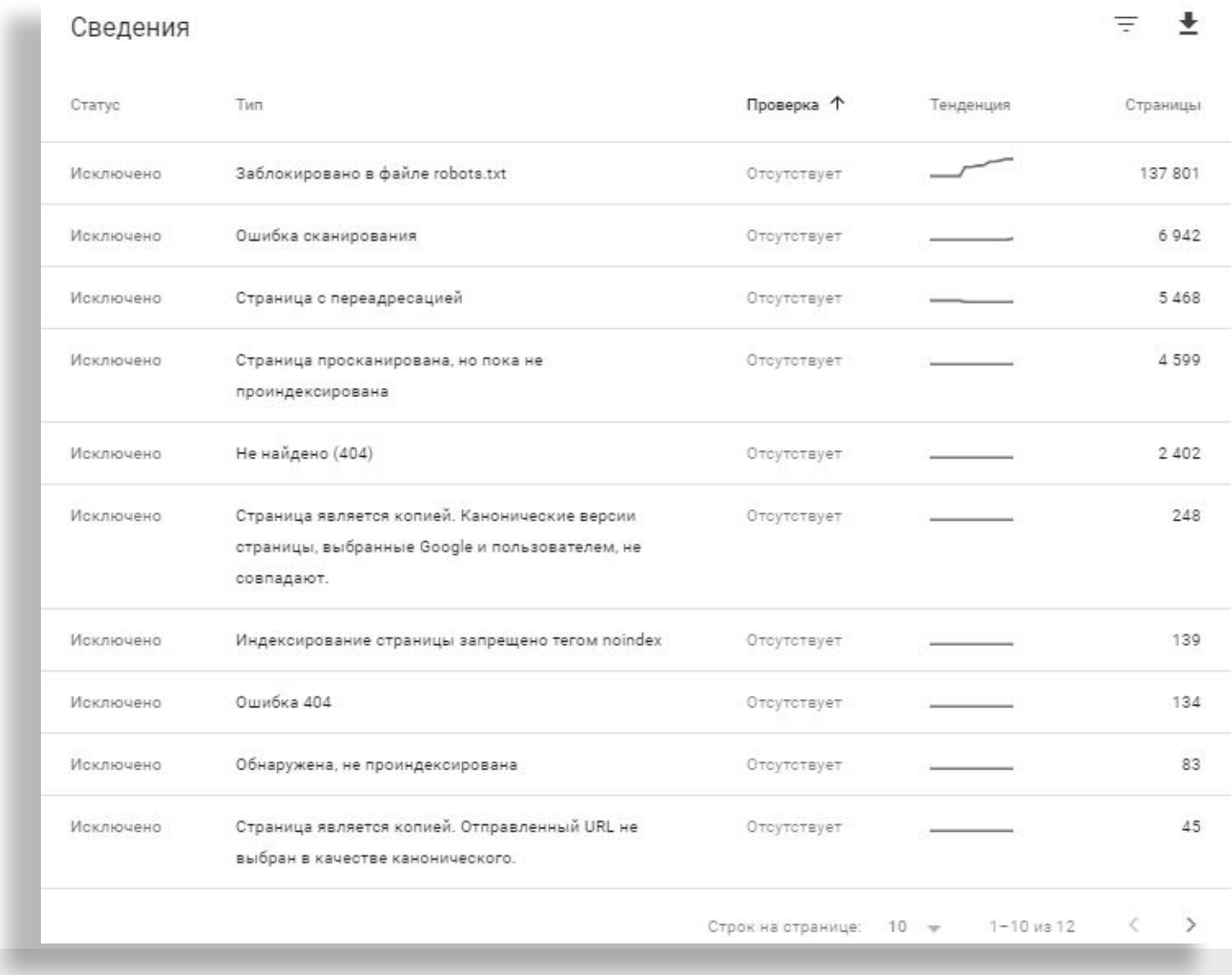
Базовые условия индексации документа, проверка индексации



Базовые условия индексации документа, проверка индексации



Базовые условия индексации документа, проверка индексации



Сведения

Статус	Тип	Проверка	Тенденция	Страницы
Исключено	Заблокировано в файле robots.txt	Отсутствует		137 801
Исключено	Ошибка сканирования	Отсутствует		6 942
Исключено	Страница с переадресацией	Отсутствует		5 468
Исключено	Страница просканирована, но пока не проиндексирована	Отсутствует		4 599
Исключено	Не найдено (404)	Отсутствует		2 402
Исключено	Страница является копией. Канонические версии страницы, выбранные Google и пользователем, не совпадают.	Отсутствует		248
Исключено	Индексирование страницы запрещено тегом noindex	Отсутствует		139
Исключено	Ошибка 404	Отсутствует		134
Исключено	Обнаружена, не проиндексирована	Отсутствует		83
Исключено	Страница является копией. Отправленный URL не выбран в качестве канонического.	Отсутствует		45

Строк на странице: 10 1–10 из 12

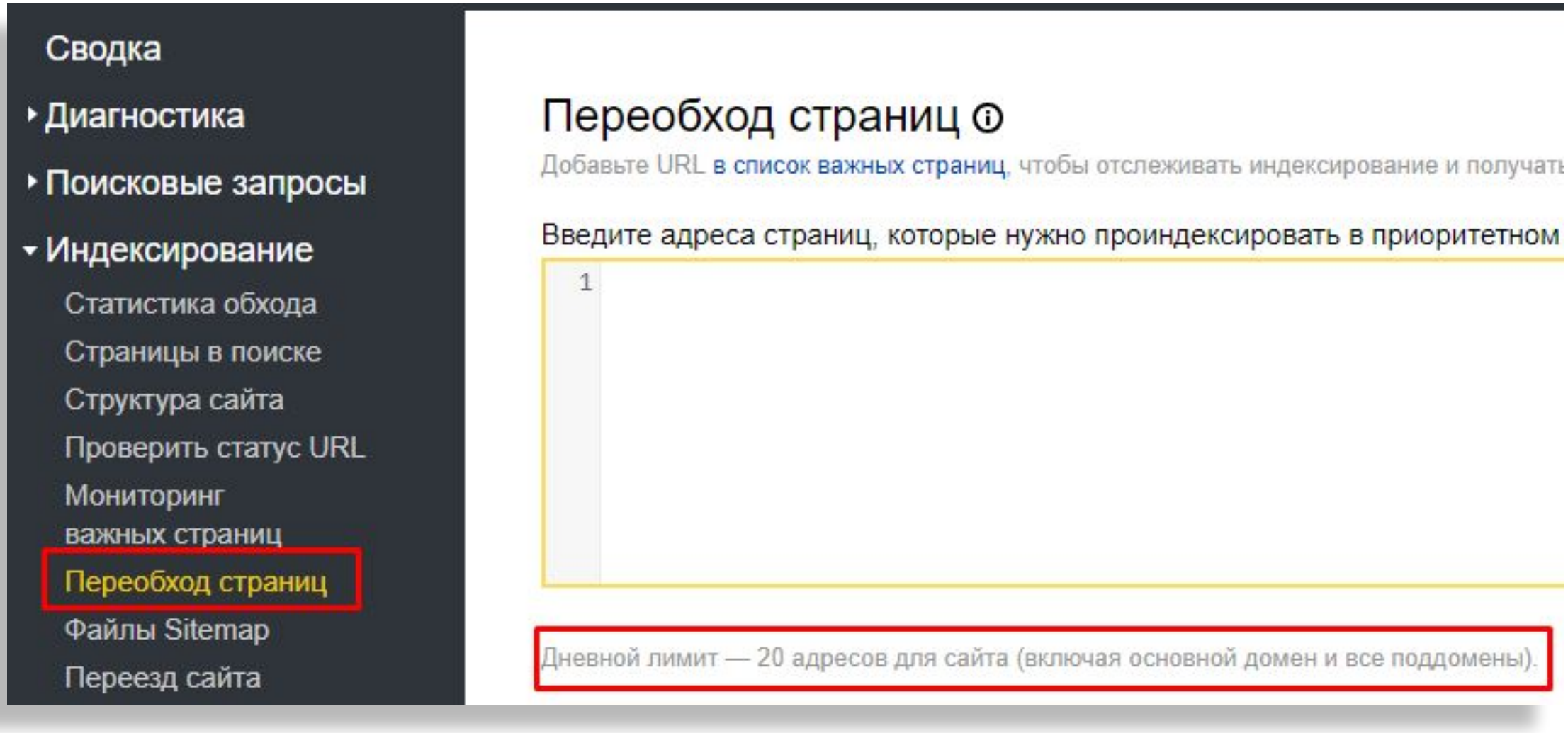
Базовые условия индексации документа, проверка индексации

Ускоряем индексацию:

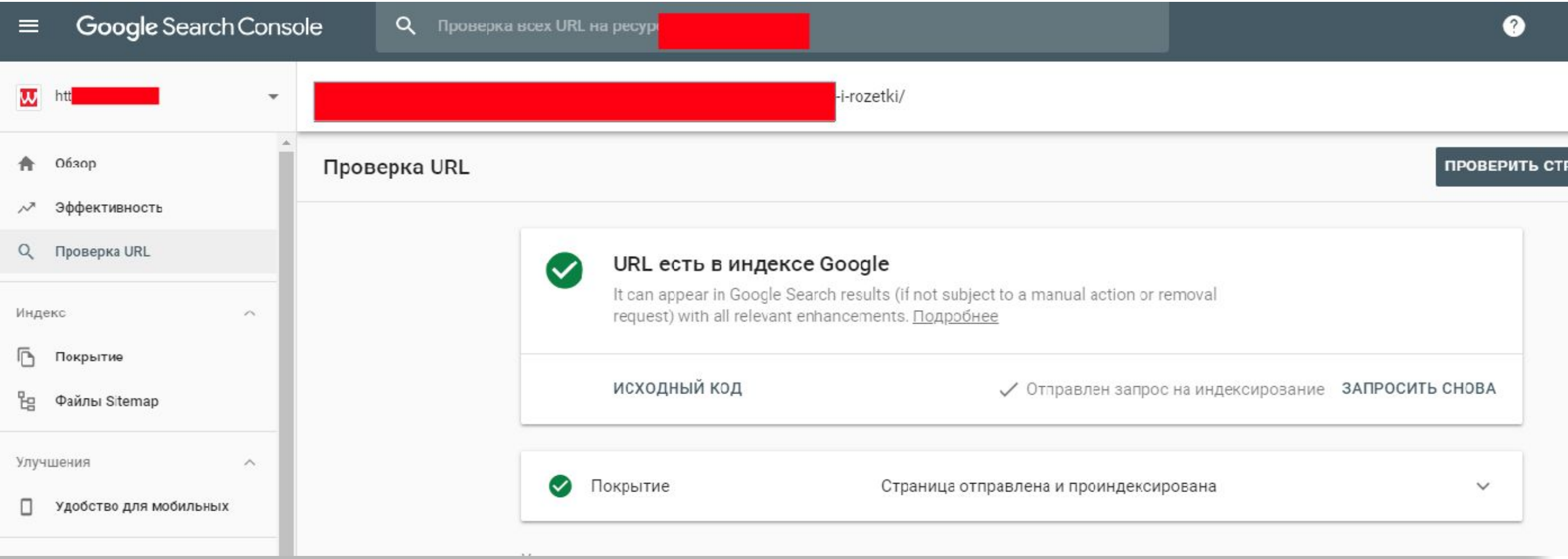
Индексирование -> Переобход страниц (в Яндекс.Вебмастер)

Сканирование -> Просмотреть как Googlebot (в Google Search Console)

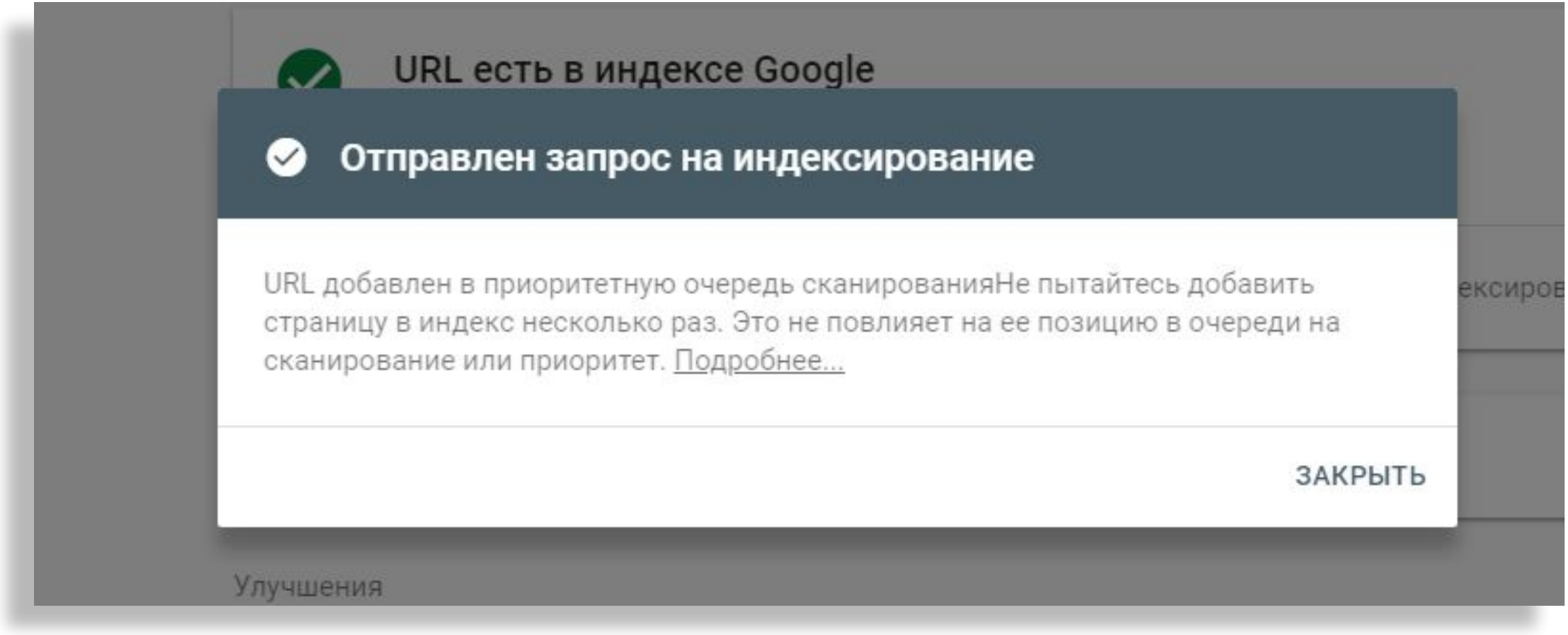
Базовые условия индексации документа, проверка индексации



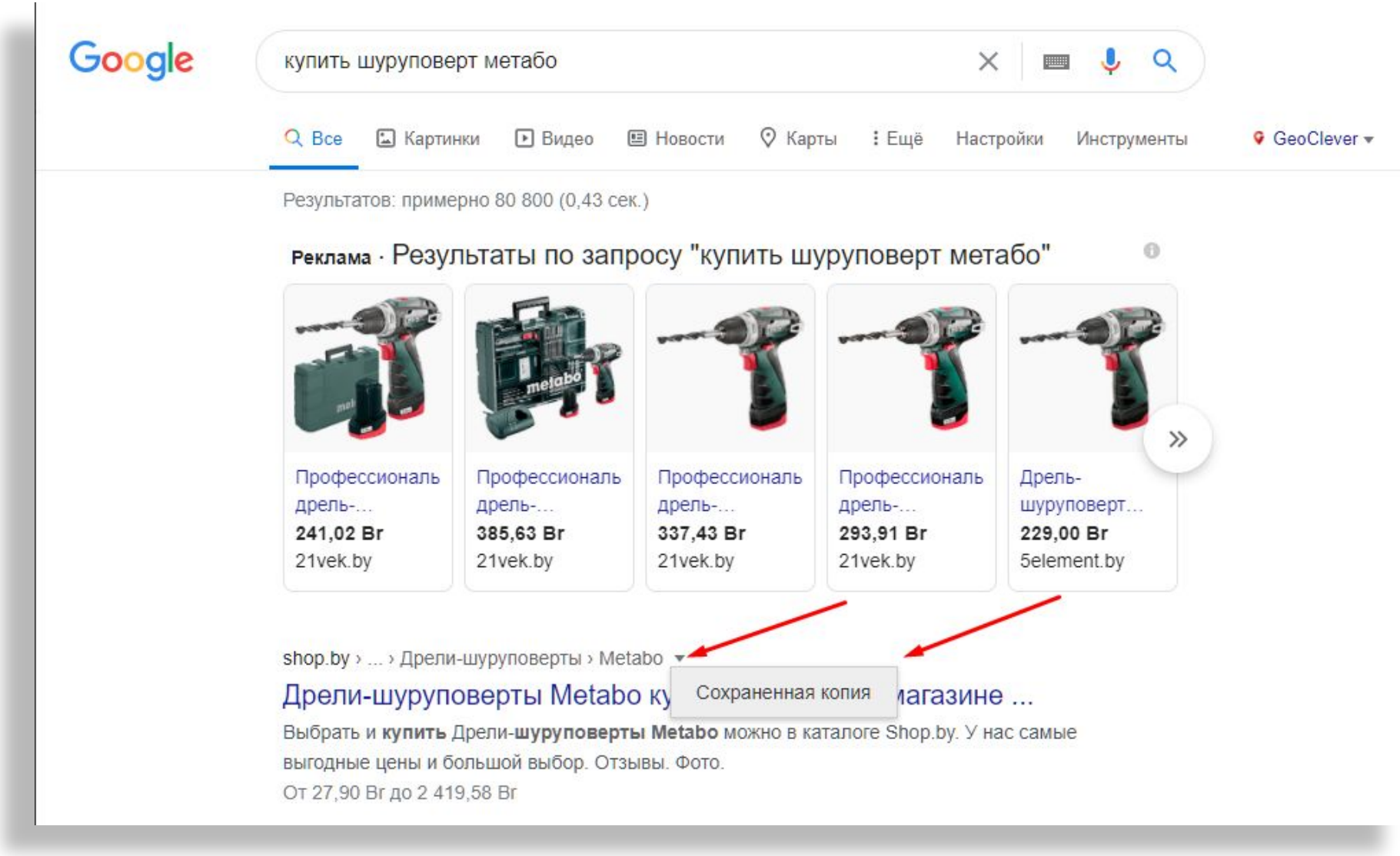
Базовые условия индексации документа, проверка индексации



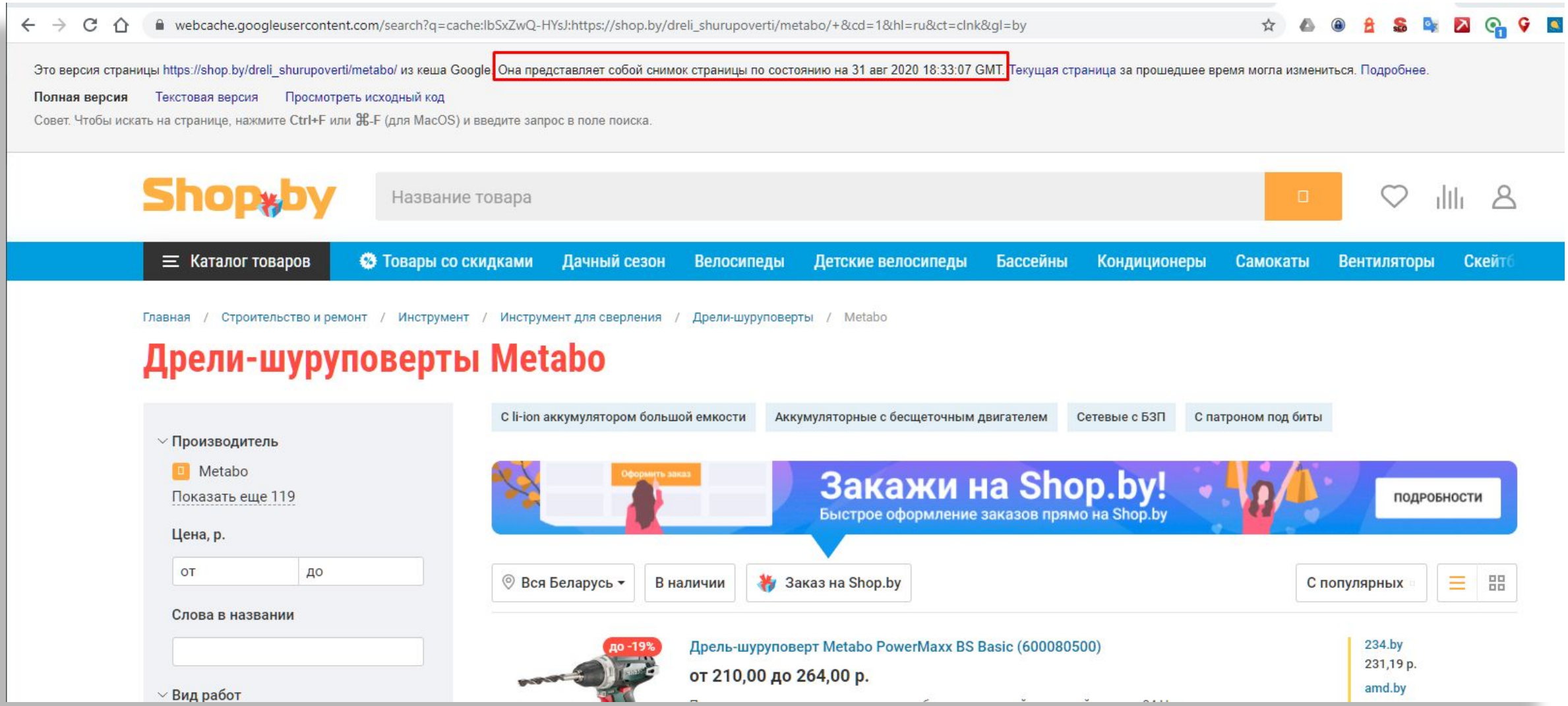
Базовые условия индексации документа, проверка индексации



Сохраненная копия



Сохраненная копия



Полные и частичные дубли: методы борьбы

Полные и частичные дубли: методы борьбы

Дубли

-это отдельные страницы сайта, контент которых полностью или частично совпадает. По сути, это копии всей страницы или ее определенной части, доступные по уникальным URL-адресам.

Дубли страниц очень опасны с точки зрения SEO. Они критично воспринимаются поисковыми системами и могут привести к серьезным потерям. Чтобы этого избежать, важно вовремя находить и удалять такие дубли.

Полные и частичные дубли: методы борьбы

Откуда могут появляться дубли:

- ✔ Автоматическая генерация дублирующих страниц движком системой управления содержимым сайта (CMS) веб-ресурса (технические дубли).
- ✔ Ошибки, допущенные вебмастерами. Например, когда один и тот же товар представлен в нескольких категориях и доступен по разным URL.
- ✔ Изменение структуры сайта, когда уже существующим страницам присваиваются новые адреса, но при этом сохраняются их дубли со старыми адресами.

Полные и частичные дубли: методы борьбы

Полные дубли - это страницы с идентичным содержимым, доступны по уникальным, неодинаковым адресам.

✔ URL-адреса страниц со слешами («/», «//», «///») и без них

site.by/catalog/page, site.by/catalog///page, site.by/catalog/page/

✔ HTTP и HTTPS страницы

https//site.by и http//site.by

✔ URL-адреса с «www» и без «www»

http//www.site.net и http//site.net.

Метод борьбы: 301 редиректы

Полные и частичные дубли: методы борьбы

http://satelit.by/catalogs/asus (6)

http://satelit.by/catalogs/asus/

http://satelit.by/catalogs////asus (7)

Полные и частичные дубли: методы борьбы

URL-адреса страниц с index.php, index.html, default.asp, default.aspx, home, home.php, main.php и т.д.:

http://site.by/index.html

http://site.by/index.php

http://site.by/home

http://site.by/catalog/index.html

http://site.by/main.php

http://site.by/index.php/category

Метод борьбы: 301 редиректы или закрытие в robots.txt

Полные и частичные дубли: методы борьбы

http://satelit.by/index.php/catalogs/asus/ (8)

Полные и частичные дубли: методы борьбы

URL-адреса страниц в верхнем и нижнем регистрах:

http://site.net/example/

http://site.net/EXAMPLE/

http://site.net/Example/

Метод борьбы: 301 редиректы

http://satelit.by/catalogs/ASUS (9)

Полные и частичные дубли: методы борьбы

Изменения в иерархической структуре URL. Например, если товар доступен по нескольким разным URL:

http://site.by/catalog/podcatalog/tovar

http://site.by/catalog/tovar

http://site.by/tovar

http://site.by/dir/tovar

Метод борьбы: ТЗ программисту – товар должен быть доступен только по 1 URL!

301 редирект для уже проиндексированных дублей (если готовы найти)

Полные и частичные дубли: методы борьбы

https://www.mitsubishielectric.kz/catalog/wall-conditioning/wall-type/series-premium/1085-premium-inverter-msz-ln60vgw/ (10)

https://www.mitsubishielectric.kz/catalog/wall-conditioning/wall-type/1085-premium-inverter-msz-ln60vgw/

Полные и частичные дубли: методы борьбы

Дополнительные параметры и метки в URL.

- ✔ Наличие меток utm, gclid, yclid и любых других динамических параметров.

http://site.by/?gclid=CjwKCAjw75HW

http://site.by/catalog/?utm_source=yandex&utm_medium=cpc

Метод борьбы: закрытие в robots.txt

Полные и частичные дубли: методы борьбы

- ✔ Первая страница пагинации каталога товаров интернет-магазина или доски объявлений, блога. Она зачастую соответствует странице категории или общей странице раздела pageall:

http://site.net/catalog

http://site.net/catalog/page1

http://site.net/catalog/?page=1

https://fd-mebel.by/gostinye/ (11)

https://fd-mebel.by/gostinye/?page=1

Метод борьбы: 301 редирект или закрытие в robots.txt

Полные и частичные дубли: методы борьбы

- ✔ Неправильные настройки 404 ошибки

http://site.net/catalog

http://site.net/catalog/asdasdadkjnwefhblsdkfmkldf

Метод борьбы: ТЗ программистам на корректную обработку несуществующих URL

http://sumki-opt.by/catalog/ (12)

http://sumki-opt.by/catalog/asdasd

Полные и частичные дубли: методы борьбы

Частичные дубли - в частично дублирующихся страницах контент одинаковый, но есть небольшие отличия в элементах.

✔ Дубли на страницах фильтров, сортировок, где есть похожее содержимое и меняется только порядок размещения. При этом текст описания и заголовки не меняются.

https://kemping.by/catalog/turizm/palatki/ (13)

https://kemping.by/catalog/turizm/palatki/?sort=PRICE&order=desc

Метод борьбы: закрытие в robots.txt

Полные и частичные дубли: методы борьбы

✔ Дубли на страницах для печати или для скачивания, основные данные которых полностью соответствуют основным страницам.

Метод борьбы: закрытие в robots.txt

https://www.21vek.by/washing_machines/iwsb51051by_indesit.html (14)

https://www.21vek.by/washing_machines/iwsb51051by_indesit.html?print

Полные и частичные дубли: методы борьбы

✔ Страницы пагинации (кроме первой)

ТЗ программистам: Уникализация title, description по шаблону, текст описания для категории должен выводиться только на первой странице (категорийная страница).

https://fd-mebel.by/gostinye/ (15)

https://fd-mebel.by/gostinye/?page=2

Полные и частичные дубли: методы борьбы

Часто решение проблемы кроется в настройке самого движка, а потому основной задачей оптимизатора является не столько устранение, сколько выявление полного списка частичных и полных дублей и постановке грамотного ТЗ исполнителю.

https://2ip.ru/cms/ - определение CMS

Служебные (мусорные) страницы

Служебные (мусорные) страницы:

Служебные\мусорные страницы

Служебные страницы:

- ✔ Корзина
- ✔ Регистрация
- ✔ Личный кабинет
- ✔ Вход в администраторскую часть
- ✔ Результаты поиска по сайту
- ✔ Технические страницы
- ✔ Тестовые страницы и т.д.

Служебные (мусорные) страницы

Что с ними делаем?

Задание для самостоятельного выполнения

Проанализируйте сайт http://it-m.by

- ✔ найдите дубли, определите их тип – полные или частичные;
- ✔ найдите служебные\мусорные страницы;
- ✔ составьте файл robots.txt в котором найденные дубли и служебные\мусорные страницы будут закрыты от индексации.

Вопросы?

Ярославцев Дмитрий
dm.yaroslavtsev@gmail.com – для ДЗ
https://www.facebook.com/yaroslavtsev.dmitriy - Для вопросов