

# 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](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](http://stalnik.by/buyer/ltrb_2016.html)

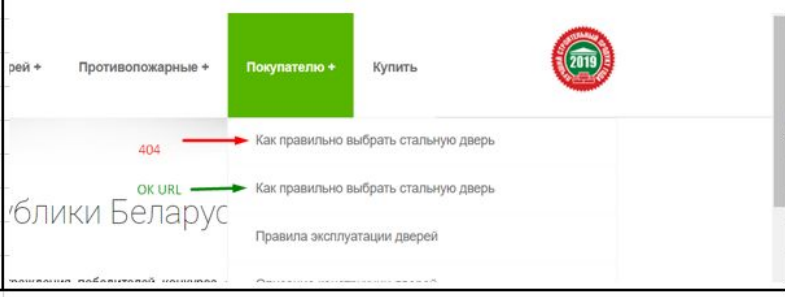
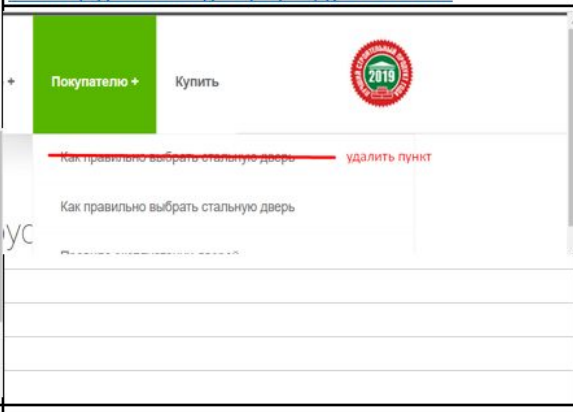
The screenshot shows the website [stalnik.by](http://stalnik.by) with a 404 error. The page title is "Конкурс «Лучшие товары Республики Беларусь»". A dropdown menu is open from the "Покупателю +" button, listing various articles and news items. The first item in the menu, "Как правильно выбрать стальную дверь", is highlighted with a red box. Below the menu is a photograph of an award ceremony where several men in suits are holding certificates.

Эту ссылку надо удалить или перелинковать на существующую страницу

# Разбор

<a href="http://stalnik.by/buyer/buyer/choice.html">http://stalnik.by/buyer/buyer/choice.html</a>	404 <u>Not Found</u>	В меню неверно указана ссылка на страницу Редикт на страницу <a href="http://stalnik.by/buyer/choice.html">http://stalnik.by/buyer/choice.html</a>
<a href="http://stalnik.by/buyer/buyer/contacts.html">http://stalnik.by/buyer/buyer/contacts.html</a>	404 <u>Not Found</u>	В меню неверно указана ссылка на страницу Редикт на страницу <a href="http://stalnik.by/buyer/contacts.html">http://stalnik.by/buyer/contacts.html</a>
<a href="http://stalnik.by/buyer/fire.html">http://stalnik.by/buyer/fire.html</a>	404 <u>Not Found</u>	Неверно указана ссылка на страницу Редикт на страницу <a href="http://stalnik.by/fire.html">http://stalnik.by/fire.html</a>
<a href="http://stalnik.by/buyer/fire2.html">http://stalnik.by/buyer/fire2.html</a>	404 <u>Not Found</u>	Неверно указана ссылка на страницу Редикт на страницу <a href="http://stalnik.by/fire2.html">http://stalnik.by/fire2.html</a>
<a href="http://stalnik.by/doors/buyer/choice.html">http://stalnik.by/doors/buyer/choice.html</a>	404 <u>Not Found</u>	Неверно указана ссылка на страницу Редикт на страницу <a href="http://stalnik.by/buyer/choice.html">http://stalnik.by/buyer/choice.html</a>
<a href="http://stalnik.by/buyer/pvokbo_2018.html">http://stalnik.by/buyer/pvokbo_2018.html</a>	404 <u>Not Found</u>	Неверно указана ссылка на страницу Редикт на страницу <a href="http://stalnik.by/buyer/pvokbo_2016.html">http://stalnik.by/buyer/pvokbo_2016.html</a>

# Разбор

A	B	C	D
<p>1 <a href="http://stalnik.by/buyer/buyer/choice.html">http://stalnik.by/buyer/buyer/choice.html</a></p> <p>2 error code: 404 (not found), linked from page(s):</p> <p>3 <a href="http://stalnik.by/buyer/ltrb_2016.html">http://stalnik.by/buyer/ltrb_2016.html</a></p> <p>4 <a href="http://stalnik.by/buyer/nkpg_2017.html">http://stalnik.by/buyer/nkpg_2017.html</a></p>	<p>на сайте в некоторых разделах, присутствует дубль ссылки на одну и ту же статью "Как выбрать входную дверь". Развернутый пункт меню "Покупателю +" на таких страницах включает в свою структуру 2 одинаковых заголовка (рис.1). Один из них имеет верную ссылку для перехода, а другой - "битую" ссылку и ведет на страницу "404 Not Found"</p> 	<p>Рекомендация по исправлению</p> <p><a href="http://stalnik.by/buyer/ltrb_2016.html">http://stalnik.by/buyer/ltrb_2016.html</a></p> <p><a href="http://stalnik.by/buyer/nkpg_2017.html">http://stalnik.by/buyer/nkpg_2017.html</a></p> 	<p>удалить лишний пункт меню с битой ссылкой</p> <p>удалить лишний пункт меню с битой ссылкой</p>

# 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 появиться вследствие изменения структуры сайта (к примеру часто встретилась ошибка в ссылках формата /buyer/buyer)

Т.е. теоретически, они могли появиться 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	<a href="http://www.stalnik.by/">http://www.stalnik.by/</a>	
2	redirected to: <a href="http://stalnik.by/">http://stalnik.by/</a>	редирект с <a href="http://www.stalnik.by/">http://www.stalnik.by/</a> на главное зеркало
3	status code: 301 (object permanently moved)	<a href="http://stalnik.by/">http://stalnik.by/</a>

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>

<http://stalnik.by/buyer/lspg 2016.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)
<a href="http://stalnik.by/buyer/buyer/choice.html">http://stalnik.by/buyer/buyer/choice.html</a>	<a href="http://stalnik.by/buyer/ltrb_2016.html">http://stalnik.by/buyer/ltrb_2016.html</a> <a href="http://stalnik.by/buyer/nkpg_2017.html">http://stalnik.by/buyer/nkpg_2017.html</a>	<a href="http://stalnik.by/buyer/choice.html">http://stalnik.by/buyer/choice.html</a>
<a href="http://stalnik.by/buyer/buyer/contacts.html">http://stalnik.by/buyer/buyer/contacts.html</a>	<a href="http://stalnik.by/buyer/choice.html">http://stalnik.by/buyer/choice.html</a> <a href="http://stalnik.by/buyer/rules.html">http://stalnik.by/buyer/rules.html</a> <a href="http://stalnik.by/buyer/construction.html">http://stalnik.by/buyer/construction.html</a> <a href="http://stalnik.by/buyer/certificate.html">http://stalnik.by/buyer/certificate.html</a> <a href="http://stalnik.by/buyer/ltrb_2016.html">http://stalnik.by/buyer/ltrb_2016.html</a> <a href="http://stalnik.by/buyer/pvokbo_2016.html">http://stalnik.by/buyer/pvokbo_2016.html</a> <a href="http://stalnik.by/buyer/nkpg_2017.html">http://stalnik.by/buyer/nkpg_2017.html</a> <a href="http://stalnik.by/buyer/lspg_2015.html">http://stalnik.by/buyer/lspg_2015.html</a> <a href="http://stalnik.by/buyer/lspg_2016.html">http://stalnik.by/buyer/lspg_2016.html</a> <a href="http://stalnik.by/buyer/lspg_2017.html">http://stalnik.by/buyer/lspg_2017.html</a> <a href="http://stalnik.by/buyer/lspg_2018.html">http://stalnik.by/buyer/lspg_2018.html</a> <a href="http://stalnik.by/buyer/contacts.html">http://stalnik.by/buyer/contacts.html</a>	<a href="http://stalnik.by/buyer/contacts.html">http://stalnik.by/buyer/contacts.html</a>
<a href="http://stalnik.by/buyer/fire.html">http://stalnik.by/buyer/fire.html</a>	<a href="http://stalnik.by/buyer/ltrb_2016.html">http://stalnik.by/buyer/ltrb_2016.html</a> <a href="http://stalnik.by/buyer/nkpg_2017.html">http://stalnik.by/buyer/nkpg_2017.html</a>	<a href="http://stalnik.by/fire.html">http://stalnik.by/fire.html</a>
<a href="http://stalnik.by/buyer/fire2.html">http://stalnik.by/buyer/fire2.html</a>	<a href="http://stalnik.by/buyer/ltrb_2016.html">http://stalnik.by/buyer/ltrb_2016.html</a> <a href="http://stalnik.by/buyer/nkpg_2017.html">http://stalnik.by/buyer/nkpg_2017.html</a>	<a href="http://stalnik.by/fire2.html">http://stalnik.by/fire2.html</a>
<a href="http://stalnik.by/doors/buyer/choice.html">http://stalnik.by/doors/buyer/choice.html</a>	<a href="http://stalnik.by/doors/stock.html">http://stalnik.by/doors/stock.html</a> <a href="http://stalnik.by/doors/flat.html">http://stalnik.by/doors/flat.html</a> <a href="http://stalnik.by/doors/house.html">http://stalnik.by/doors/house.html</a> <a href="http://stalnik.by/doors/house-by-name_two.html">http://stalnik.by/doors/house-by-name_two.html</a> <a href="http://stalnik.by/doors/house-by-name.html">http://stalnik.by/doors/house-by-name.html</a>	<a href="http://stalnik.by/buyer/choice.html">http://stalnik.by/buyer/choice.html</a>
<a href="http://stalnik.by/doors/stock/buyer/choice.html">http://stalnik.by/doors/stock/buyer/choice.html</a>	<a href="http://stalnik.by/doors/stock/optima.html">http://stalnik.by/doors/stock/optima.html</a> <a href="http://stalnik.by/doors/stock/ritm.html">http://stalnik.by/doors/stock/ritm.html</a> <a href="http://stalnik.by/doors/stock/vector.html">http://stalnik.by/doors/stock/vector.html</a> <a href="http://stalnik.by/doors/stock/logika.html">http://stalnik.by/doors/stock/logika.html</a> <a href="http://stalnik.by/doors/stock/universal.html">http://stalnik.by/doors/stock/universal.html</a>	<a href="http://stalnik.by/buyer/choice.html">http://stalnik.by/buyer/choice.html</a>

# Курс SEO-практик

---

## Управление индексацией сайта. Дубли и служебные страницы

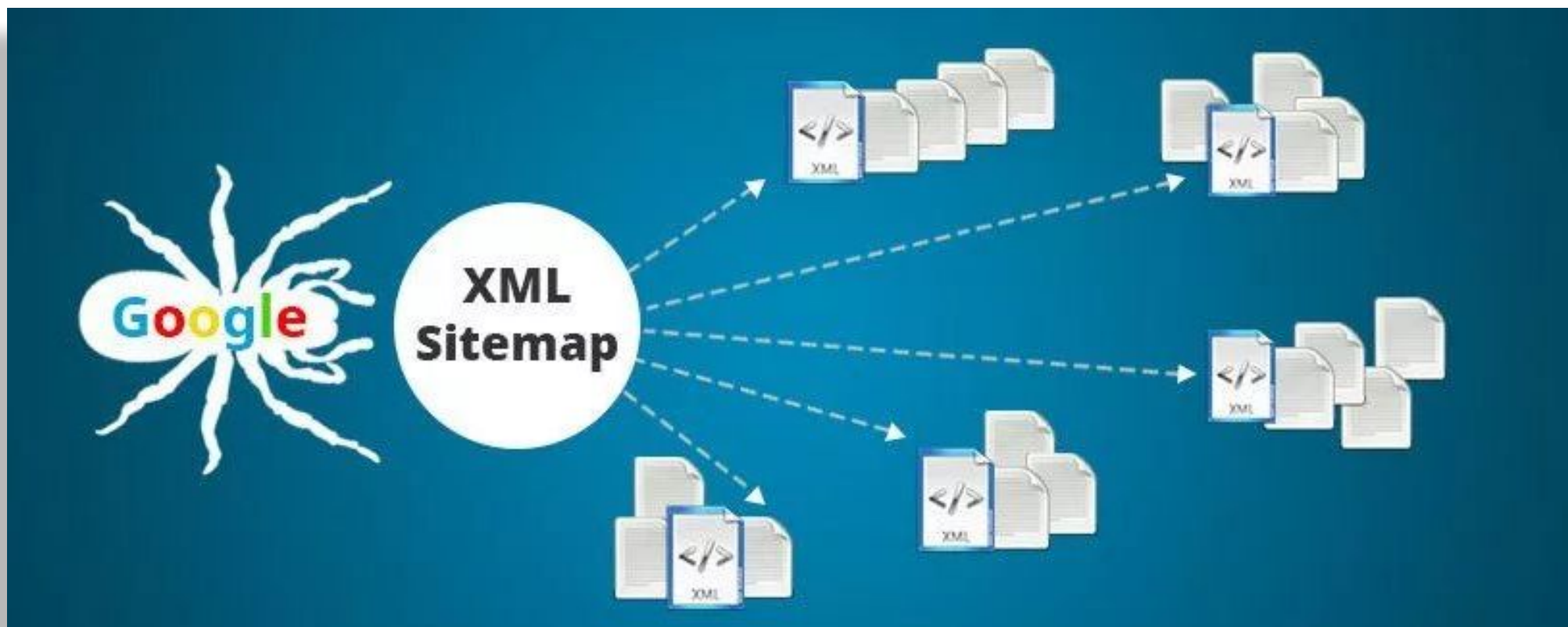
Модуль 7



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

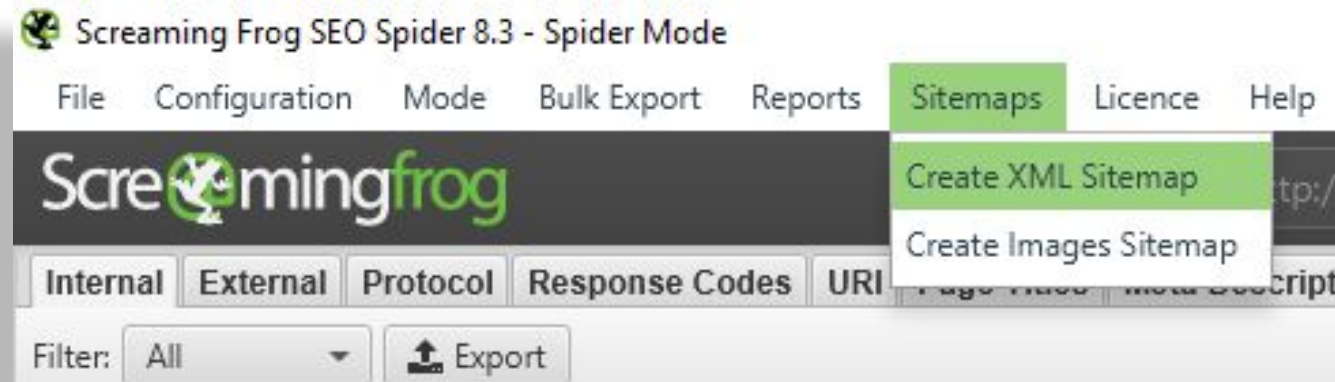
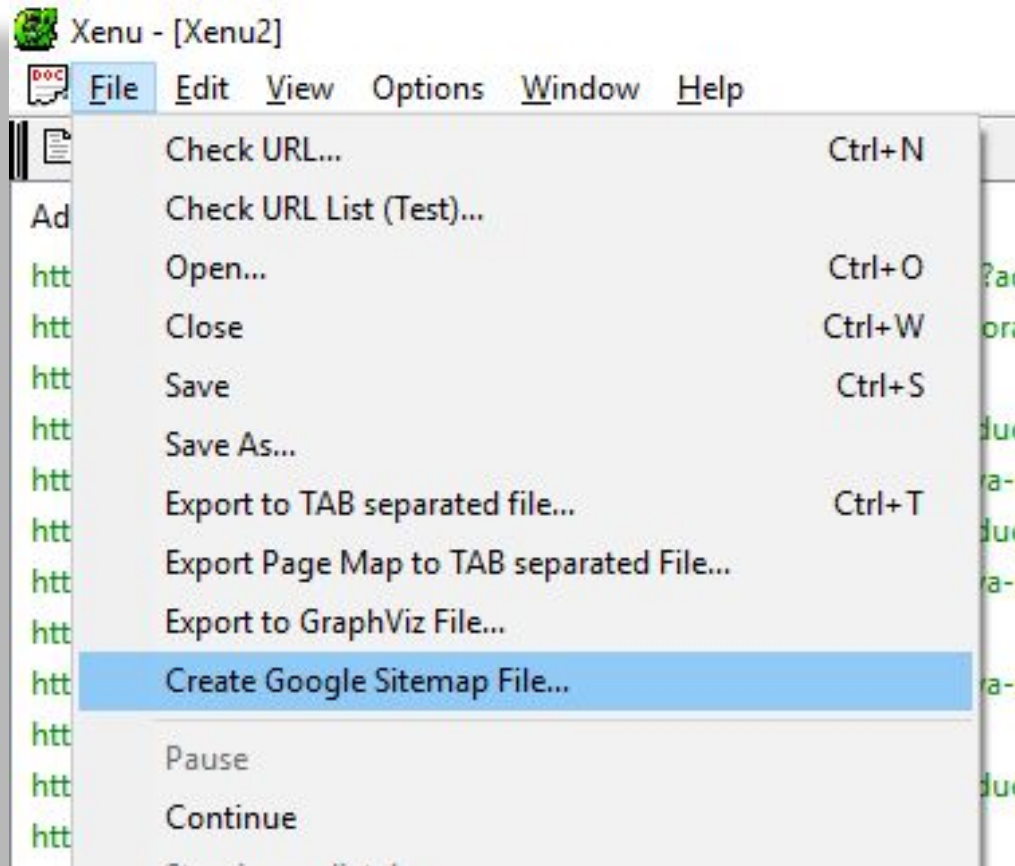
## Как создать Sitemap.xml

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

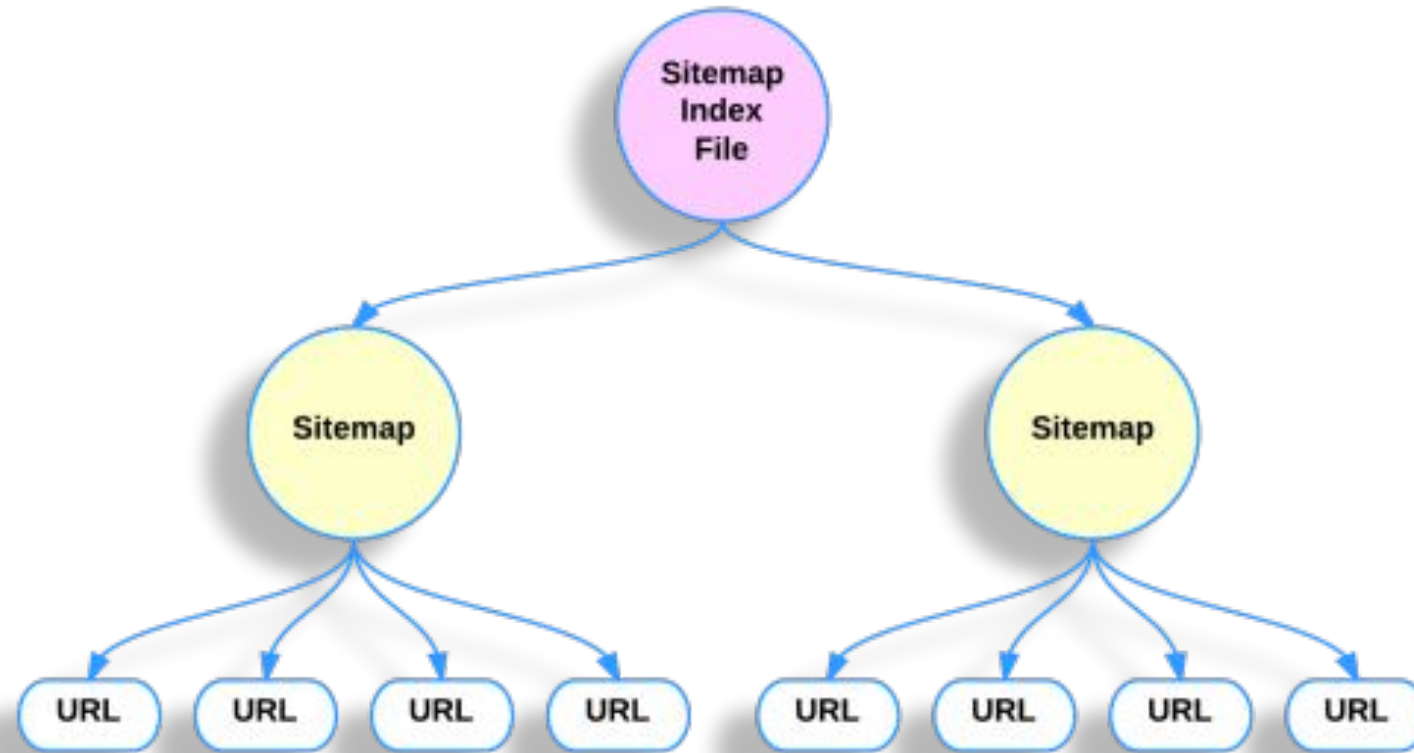
<https://www.termebel.by/sitemap.xml> (1)

# 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

- Качество сайта •
- Диагностика
- Поисковые запросы •
- ▾ Индексирование •
  - Статистика обхода
  - Страницы в поиске
  - Структура сайта
  - Проверить статус URL
  - Мониторинг важных страниц
  - Переобход страниц
  - Файлы Sitemap
  - Обход по счётчикам •
  - Переезд сайта
  - Скорость обхода
- Ссылки
- Информация о сайте
- Турбо-страницы •
- Инструменты •
- Настройки •
- Полезные сервисы

Добавить файл Sitemap ⓘ

Используемые файлы Sitemap

Источник (можно ↻ отправить на переобход еще 10)

Источник	Статус	Последняя загрузка	Число ссылок в файле
https://[REDACTED] sitemap.xml ×	↻ ок	03.04.2019, 4:17	6
- https://[REDACTED] sitemap_files.xml	↻ ок	03.04.2019, 9:59	8
- https://[REDACTED] sitemap_iblock_12.part1.xml	↻ 4 ошибки	29.03.2019, 1:50	0
- https://[REDACTED] sitemap_iblock_3.xml	↻ ок	02.04.2019, 19:47	17
- https://[REDACTED] sitemap_iblock_1.xml	↻ ок	03.04.2019, 11:52	33
- https://[REDACTED] sitemap_iblock_12.xml	↻ ок	25.03.2019, 15:52	2 334
- https://[REDACTED] sitemap_iblock_2.xml	↻ ок	01.04.2019, 7:47	6

Найдены в robots.txt

https://[REDACTED] sitemap.xml	↻ ок	03.04.2019, 4:17	6
- https://[REDACTED] sitemap_files.xml	↻ ок	03.04.2019, 9:59	8
- https://[REDACTED] sitemap_iblock_12.part1.xml	↻ 4 ошибки	29.03.2019, 1:50	0
- https://[REDACTED] sitemap_iblock_3.xml	↻ ок	02.04.2019, 19:47	17
- https://[REDACTED] sitemap_iblock_1.xml	↻ ок	03.04.2019, 11:52	33
- https://[REDACTED] sitemap_iblock_12.xml	↻ ок	25.03.2019, 15:52	2 334
- https://[REDACTED] sitemap_iblock_2.xml	↻ ок	01.04.2019, 7:47	6

# Sitemap.xml наиболее частые ошибки

The screenshot shows the Google Search Console interface for a website. The main heading is "Файлы Sitemap". Below it, there is a form to "Добавьте файл Sitemap" with a text input field containing "https://[redacted]/ URL файла Sitemap" and a button labeled "ОТПРАВИТЬ".

Below the form is a table titled "Файлы Sitemap на проверке". The table has the following columns: Sitemap, Тип, Отправлено, Дата последней обработки, Статус, and Количество выявленных URL.

Sitemap	Тип	Отправлено	Дата последней обработки	Статус	Количество выявленных URL
/sitemap.xml	Sitemap	5 апр. 2019 г.	6 февр. 2019 г.	1 ошибка	5 000
/index.php?route=feed/blog_sitemap	Неизвестно	24 мар. 2019 г.		Не получено	0

# 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](https://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](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

```
User-Agent: *  
Disallow: /wp-admin/  
Disallow: /wp-includes/  
Disallow: /wp-content/  
Disallow: /wp-  
Disallow: /*.css$  
  
user-agent: Googlebot-1  
Disallow:  
  
user-agent: Mediapartner  
Disallow:
```



# 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

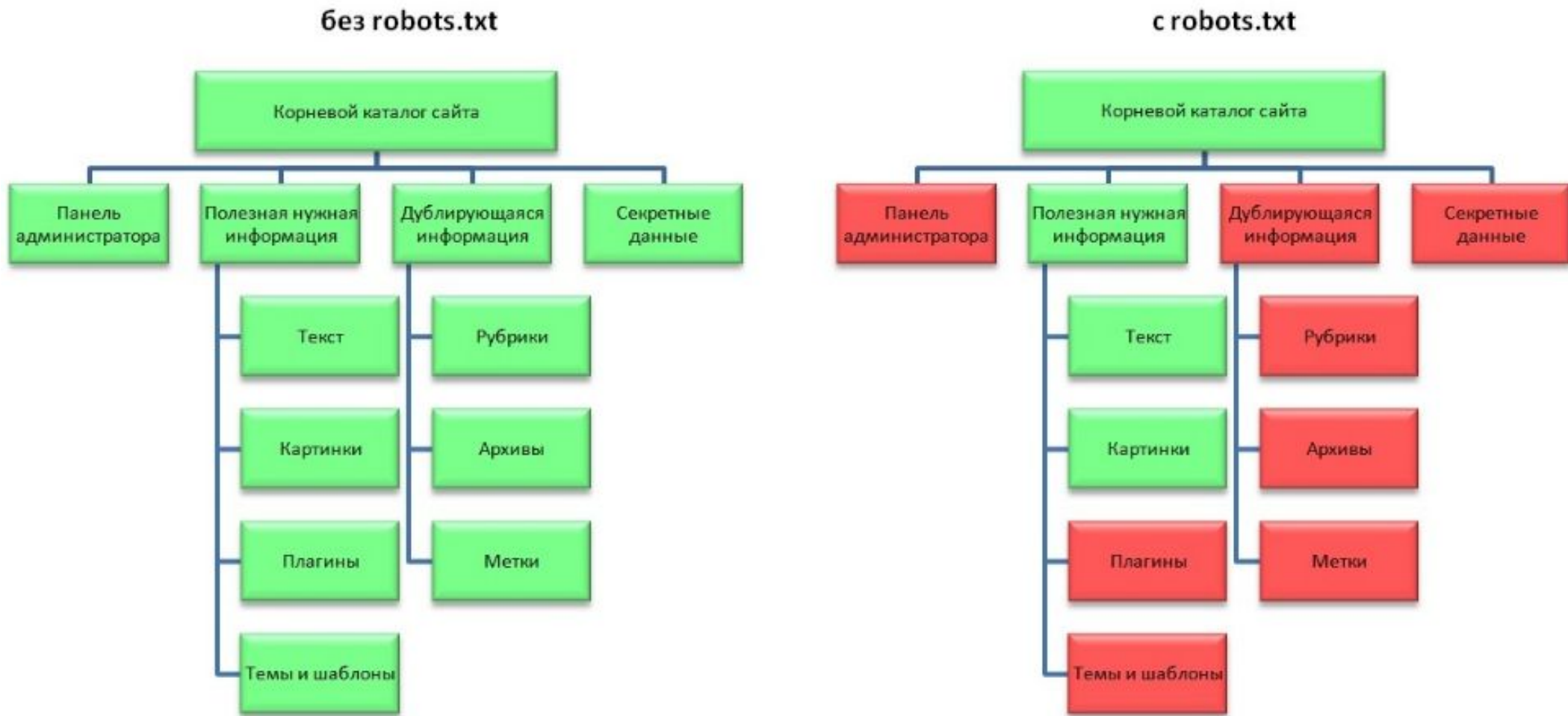
Например, мы не хотим, чтобы роботы поисковых систем посещали:

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

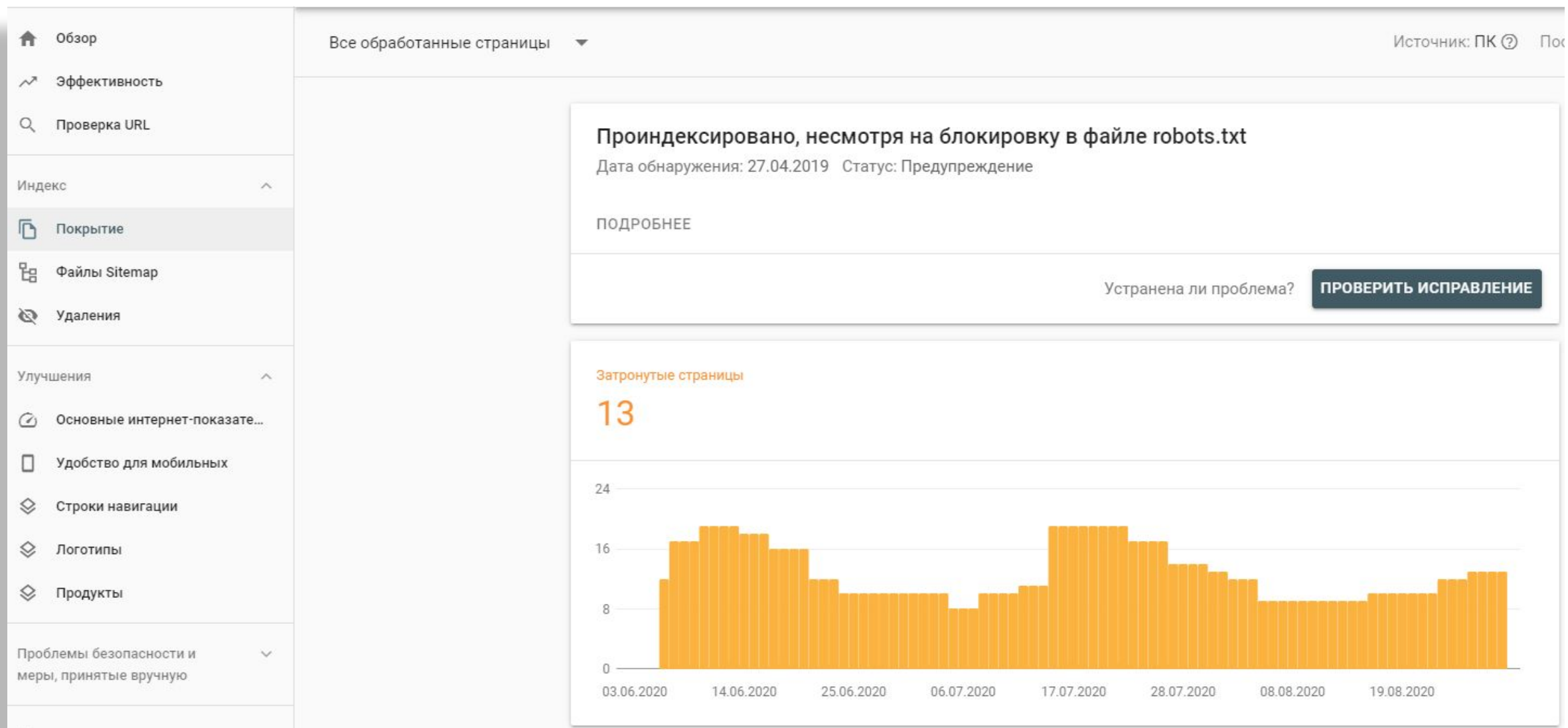
**Важно понимать, что закрытие страницы не является 100% гарантией того, что робот ее не проиндексирует!**

# Robots.txt директивы и их использование

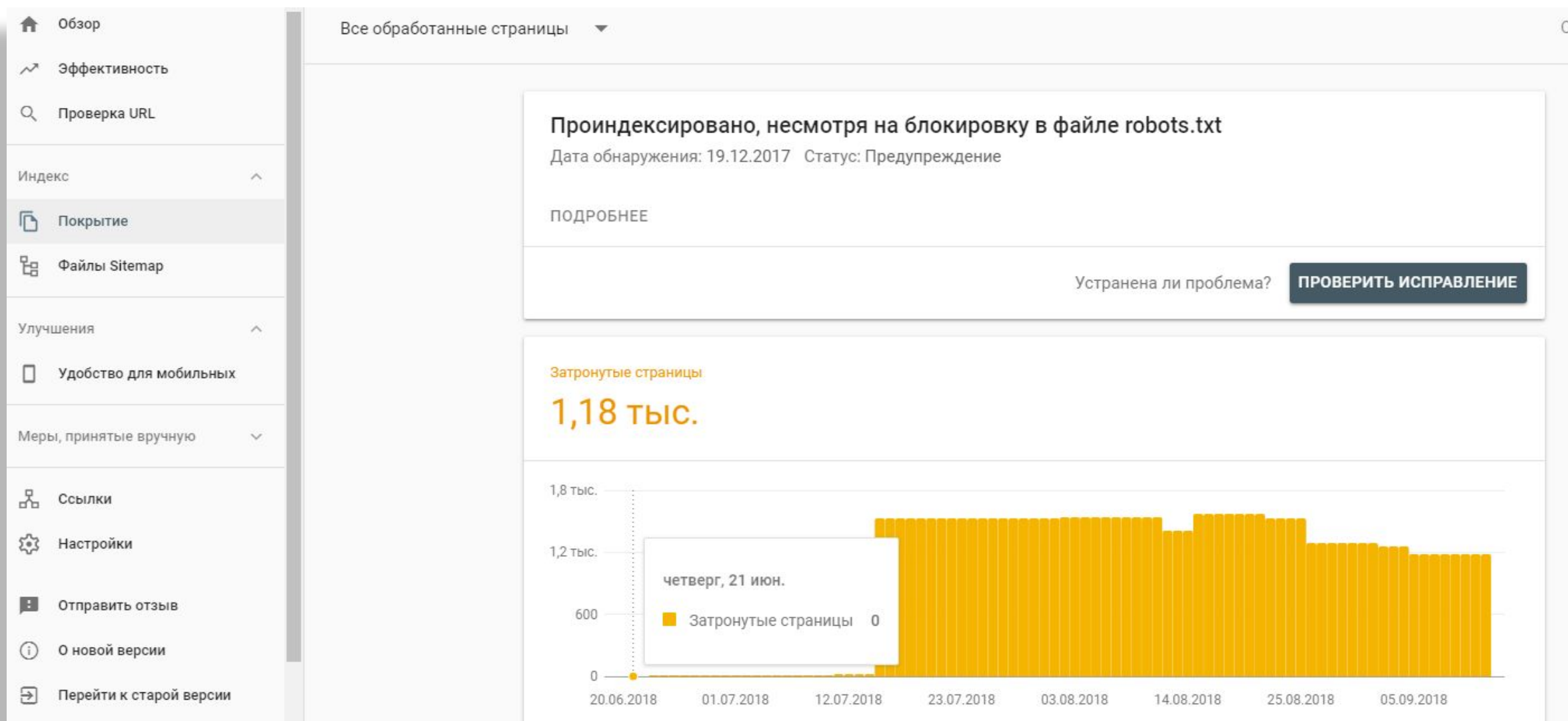
Пример того, что индексируется поисковиками:



# 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)  
<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 обрабатывать 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

## Анализ robots.txt ⓘ

Проверяемый сайт

http://fastgreen.by/

```
1 User-agent: *
2 Allow: /*js
3 Allow: /*css
4 Allow: /*png
5 Allow: /*jpg
6 Allow: /*gif
7 Allow: /wp-content/themes/
8 Allow: /wp-includes/js/wp-embed min is?ver=4.4.7
```

Проверить



# 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
<a href="https://linenmill.by/kontraktnyj-zakaz/">https://linenmill.by/kontraktnyj-zakaz/</a>	<a href="#">/kontraktnyj-zakaz/\$</a>
<a href="https://linenmill.by/author/vova/">https://linenmill.by/author/vova/</a>	<a href="#">/author/vova/\$</a>
<a href="https://linenmill.by/author/zenya/">https://linenmill.by/author/zenya/</a>	<a href="#">/author/zenya/\$</a>

# 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 вебмастеров Яндекса и Google

✓ Запросы с использованием операторов

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

`url:site.by/*` - Яндекс для сайта

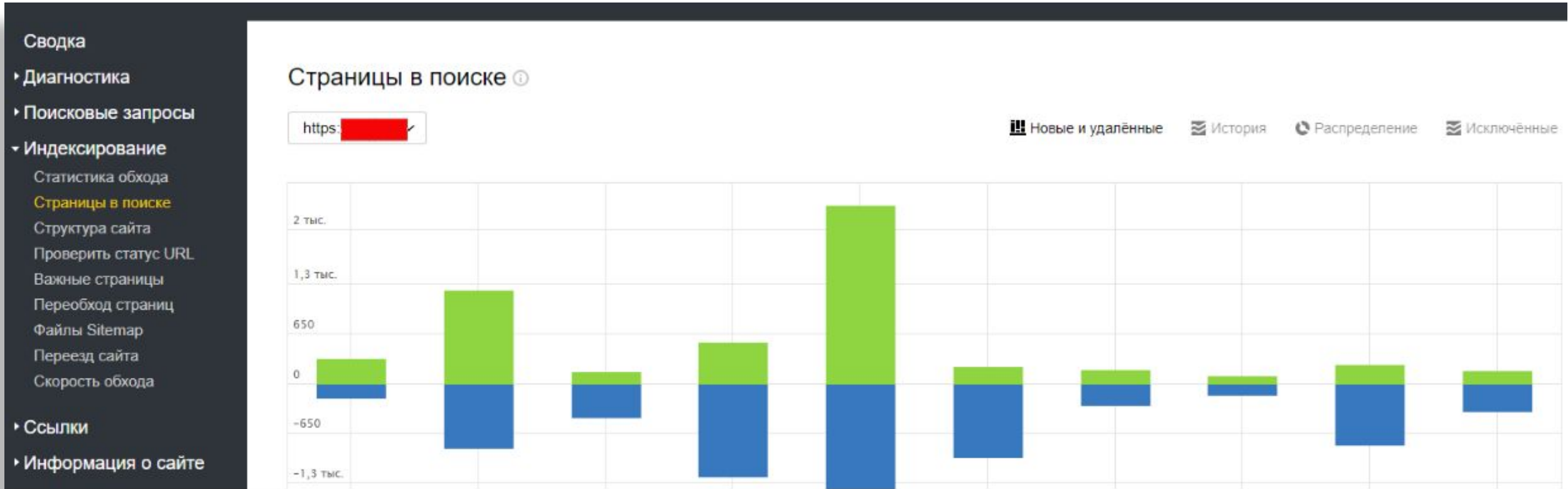
~~`info:https://site.by/catalog/page1.html` - Google для страницы~~

`site:site.by` - Google для сайта

✓ Плагины для браузера, например, RDS bar



# Базовые условия индексации документа, проверка индексации



# Базовые условия индексации документа, проверка индексации

<b>Сводка</b> ▶ Диагностика ▶ Поисквые запросы ▶ Индексирование Статистика обхода Страницы в поиске <b>Структура сайта</b> Проверить статус URL Важные страницы Переобход страниц Файлы Sitemap Перезд сайта	<b>Структура сайта</b> ⓘ			
	<b>Раздел</b>	<b>Загружено</b>	<b>В поиске</b>	<b>Доля загруженных, %</b>
	https ██████████	20 696	9 415	100
	└ /catalog	20 186	9 135	98
	└ /brands	430	258	2
Пользовательские разделы (0 из 5)				
<input type="button" value="Добавить раздел"/>				




# Базовые условия индексации документа, проверка индексации

- Сводка
- Диагностика
- Поисквые запросы
- ▾ Индексирование
  - Статистика обхода
  - Страницы в поиске
  - Структура сайта
  - Проверить статус URL**
  - Важные страницы
  - Переобход страниц
  - Файлы Sitemap
  - Переезд сайта
  - Скорость обхода
- Ссылки
- Информация о сайте
- Турбо-страницы 🚀
- Инструменты

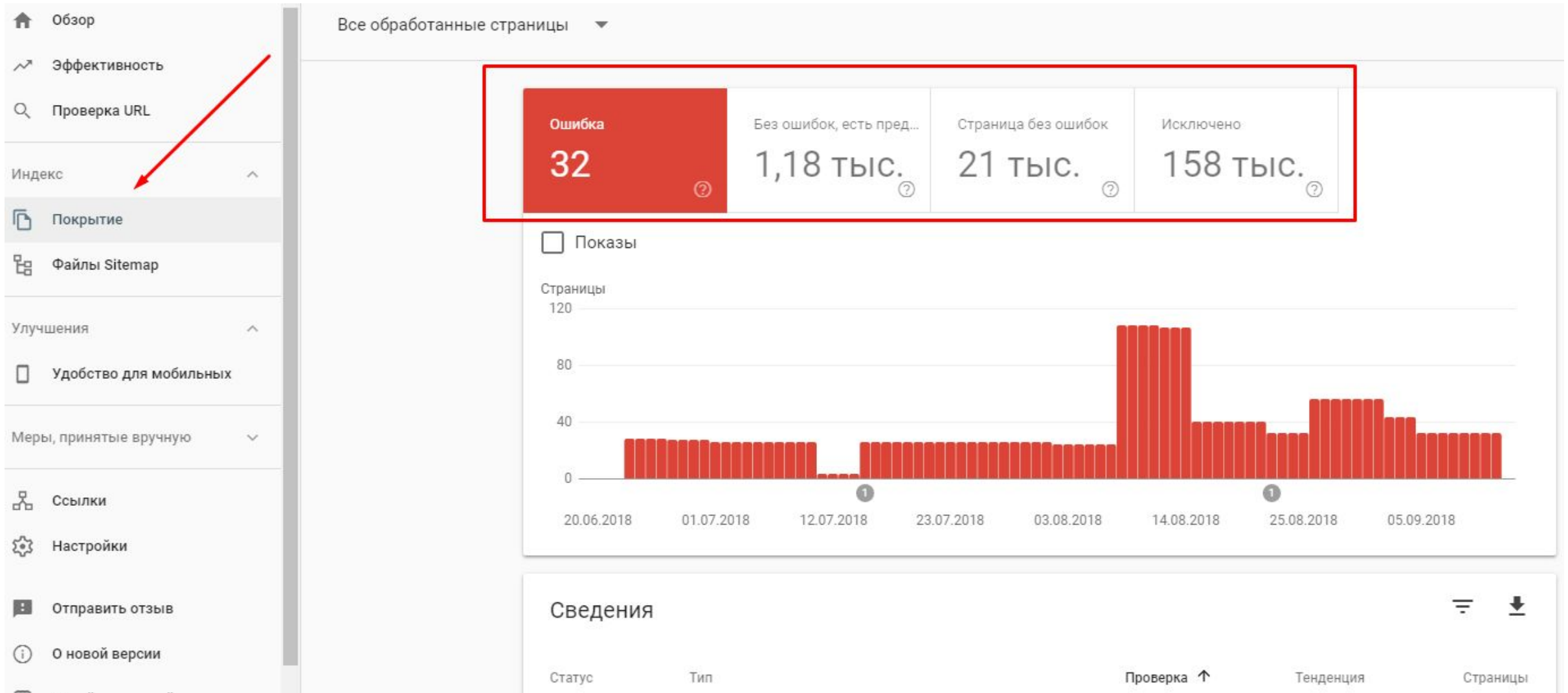
## Проверить статус URL ⓘ

Добавьте URL в список [важных страниц](#), чтобы отслеживать индексирование и получать уведомления об изменении его статуса в поиске. [Подробнее](#)











### Статус проверки ▾ Страница и результат проверки

 Проверено	<a href="#">https://[redacted]int-kle...</a>	<a href="#">Подробнее</a>	<a href="#">Отслеживать</a>	
30.11.2017 – 14:34	Имп. блок пит. 120Вт, 5А, 85...264VAC (120...375VDC) / 24VDC, DIN35, винт. клеммы, ал. корпус			
	Страница обходится роботом, но отсутствует в поиске.			
	Статус в поисковой базе: 200 ОК			
	Версия страницы: 26.10.2017 – 08:51			
	Последний обход: 200 ОК			
	Версия страницы: 26.10.2017 – 08:51			
	Страница исключена из поиска в результате работы <a href="#">специального алгоритма</a> . Если в будущем алгоритм сочтёт страницу достаточно релевантной запросам пользователей, она появится в поиске автоматически.			
	<a href="#">Ознакомиться с рекомендациями</a>			
 Проверено	<a href="#">https://[redacted]spayach...</a>	<a href="#">Подробнее</a>	<a href="#">Отслеживать</a>	
02.08.2017 – 11:25	СИСТЕМЫ ДЛЯ ПРОКЛАДКИ КАБЕЛЕВ. Точка учета энергии: #31.000. #31.000. Офис: Бунд-#31. #31.			

# Базовые условия индексации документа, проверка индексации



# Базовые условия индексации документа, проверка индексации

Сведения <span style="float: right;">☰ ⬇</span>				
Статус	Тип	Проверка ↑	Тенденция	Страницы
Исключено	Заблокировано в файле robots.txt	Отсутствует		137 801
Исключено	Ошибка сканирования	Отсутствует		6 942
Исключено	Страница с перенадресацией	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

Сводка

▸ Диагностика

▸ Поисквые запросы

▾ Индексирование

Статистика обхода

Страницы в поиске

Структура сайта

Проверить статус URL

Мониторинг

важных страниц

**Переобход страниц**

Файлы Sitemap

Переезд сайта

## Переобход страниц ⓘ

Добавьте URL в список **важных страниц**, чтобы отслеживать индексирование и получать

Введите адреса страниц, которые нужно проиндексировать в приоритетном

1

Дневной лимит — 20 адресов для сайта (включая основной домен и все поддомены).

# Базовые условия индексации документа, проверка индексации

The screenshot displays the Google Search Console interface. At the top, the search bar contains the text "Проверка всех URL на ресурсе" followed by a redacted domain. The left sidebar shows navigation options: "Обзор", "Эффективность", "Проверка URL", "Индекс", "Покрытие", "Файлы Sitemap", "Улучшения", and "Удобство для мобильных". The main content area is titled "Проверка URL" and shows two successful check results for the URL "http://[redacted]-i-rozetki/".

**Проверка URL** ПРОВЕРИТЬ СТРАНИЦУ

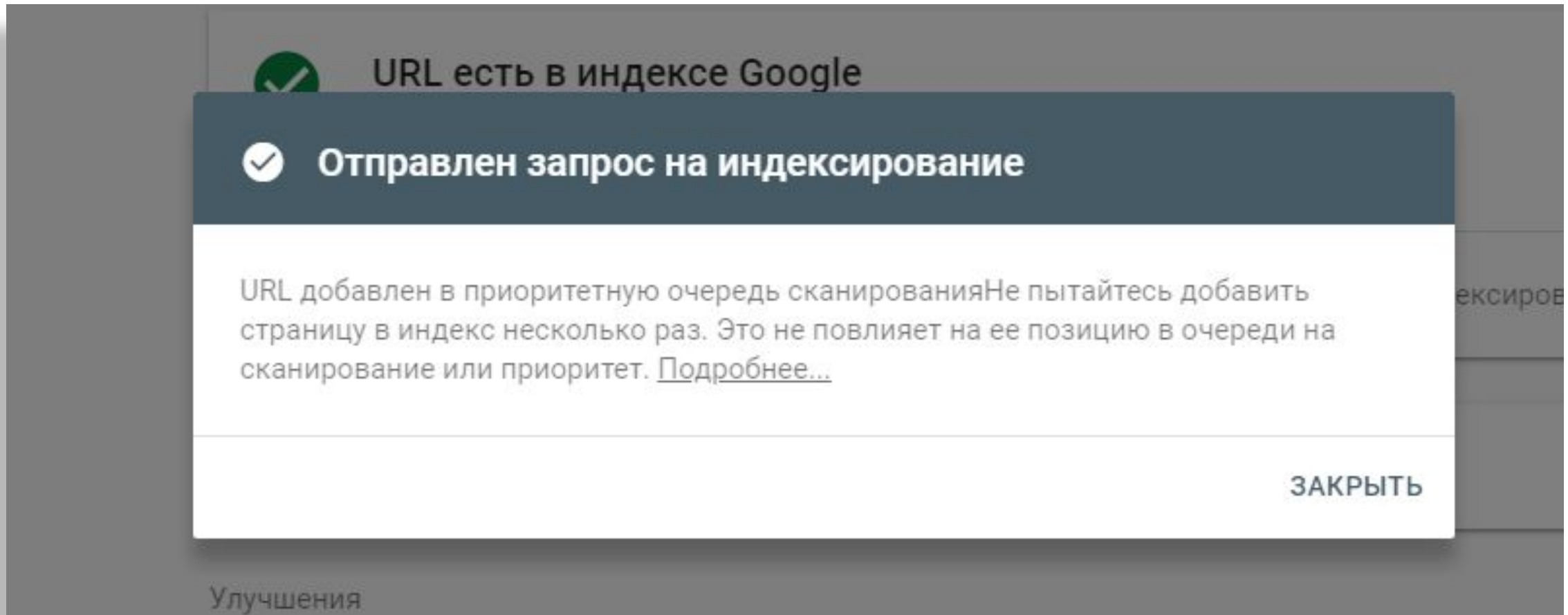
**✓ URL есть в индексе Google**  
It can appear in Google Search results (if not subject to a manual action or removal request) with all relevant enhancements. [Подробнее](#)

ИСХОДНЫЙ КОД ✓ Отправлен запрос на индексирование **ЗАПРОСИТЬ СНОВА**

**✓ Покрытие** Страница отправлена и проиндексирована



# Базовые условия индексации документа, проверка индексации



# Сохраненная копия

Google

купить шуруповерт метабо

Все Картинки Видео Новости Карты Ещё Настройки Инструменты GeoClever

Результатов: примерно 80 800 (0,43 сек.)

Реклама · Результаты по запросу "купить шуруповерт метабо"

Профессиональ дрель-... <b>241,02 Br</b> 21vek.by	Профессиональ дрель-... <b>385,63 Br</b> 21vek.by	Профессиональ дрель-... <b>337,43 Br</b> 21vek.by	Профессиональ дрель-... <b>293,91 Br</b> 21vek.by	Дрель- шуруповерт... <b>229,00 Br</b> 5element.by

shop.by > ... > Дрели-шуруповерты > Metabo

**Дрели-шуруповерты Metabo** ку Сохраненная копия магазине ...

Выбрать и **купить Дрели-шуруповерты Metabo** можно в каталоге Shop.by. У нас самые выгодные цены и большой выбор. Отзывы. Фото.

От 27,90 Br до 2 419,58 Br



# 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](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/asdasdadkijnwefhblsdkfmklidf>*

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

[https://www.21vek.by/washing\\_machines/iwsb51051by\\_indesit.html?print](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.

## 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](mailto:dm.yaroslavtsev@gmail.com) – для ДЗ

<https://www.facebook.com/yaroslavtsev.dmitriy> - Для вопросов