Техническое SEO открытых архивов на платформе DSpace

Александр Ефимов
Уральский федеральный университет

alexander.efimov@urfu.ru

Базовые вещи

- robots.txt
- sitemap
- page title
- favicon
- и пр.

Пример реального robots.txt старой версии Dspace

```
====
    The contents of this file are subject to the license and copyright
    detailed in the LICENSE and NOTICE files at the root of the source
    tree and available online at

    http://www.dspace.org/license/
====
User-agent: *

# Uncomment the following line ONLY if sitemaps.org or HTML sitemaps are used
# and you have verified that your site is being indexed correctly.
# Disallow: /browse

# You also may wish to disallow access to the following paths, in order
# to stop web spiders from accessing user-based content:
# Disallow: /advanced-search
# Disallow: /contact
# Disallow: /feedback
# Disallow: /forgot
# Disallow: /login
# Disallow: /register
# Disallow: /search
```

Пример стандартного robots.txt «свежей» версии DSpace

```
# The FULL URL to the DSpace sitemaps
# The http://demo.dspace.org/xmlui will be auto-filled with the value in dspace.cfg
# XML sitemap is listed first as it is preferred by most search engines
Sitemap: http://demo.dspace.org/xmlui/sitemap
Sitemap: http://demo.dspace.org/xmlui/htmlmap

##########################
# Default Access Group
# (NOTE: blank lines are not allowable in a group record)
##########################
User-agent: *
# Disable access to Discovery search and filters
Disallow: /discover
Disallow: /search-filter

#
# Optionally uncomment the following line ONLY if sitemaps are working
# and you have verified that your site is being indexed correctly.
# Disallow: /browse
# Disallow: /handle/10673/*/browse
#
# If you have configured DSpace (Solr-based) Statistics to be publicly
# accessible, then you may not want this content to be indexed
# Disallow: /statistics
#
# You also may wish to disallow access to the following paths, in order
# to stop web spiders from accessing user-based content
# Disallow: /contact
# Disallow: /feedback
# Disallow: /forgot
# Disallow: /login
# Disallow: /register


##############################
# Section for misbehaving bots
# The following directives to block specific robots were borrowed from Wikipedia's robots.txt
##############################

# advertising-related bots:
User-agent: Mediapartners-Google*
Disallow: /

# Crawlers that are kind enough to obey, but which we'd rather not have
# unless they're feeding search engines.
User-agent: UbiCrawler
Disallow: /
```

А вот как сделал бы SEOшник…

```
Host: elar.rsvpu.ru

Sitemap: http://elar.rsvpu.ru/sitemap
Sitemap: http://elar.rsvpu.ru/htmlmap

User-agent: *
Allow: /
Disallow: /xmlui/

User-agent: Yandex
Allow: /
Disallow: /oai/
Disallow: /xmlui/

User-agent: Baiduspider
Allow: /
Disallow: /xmlui/
```

- Директива host
- Директивы sitemap
- Специфические правила для конкретных поисковых систем

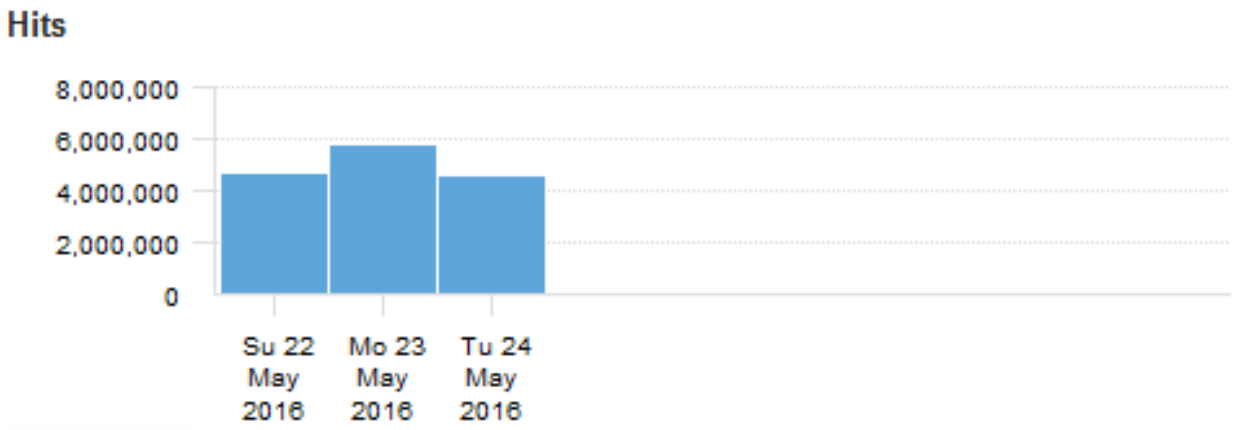
!NB

Харвестеры не уважают (do not respect) наши robots.txt, так что увлекаться «баном» конкретных юзер-агентов не стоит.

Одна из самых серьезных «нагрузок» на наши сервера создавалась сервисом scrapinghub.com, который позволял использовать любой user-agent.

Веселая картинка

Days



1 - 3 of 3

Rows

	↑ Day	Hits	Page views	Visitors	Size	Sessions	Session duration
1	22/May/2016	4,596,905	4,585,034	7,912	113.58 G	10,621	90d 03:33:37
2	23/May/2016	5,676,582	5,660,831	9,499	145.08 G	12,257	96d 19:03:37
3	24/May/2016	4,460,919	4,445,358	9,420	130.79 G	12,282	105d 02:22:18
	Total	**14,734,406**	**14,691,223**	–	**389.45 G**	–	**292d 00:59:32**

Стандартные sitemap и htmlmap

- Их почти никто не создаёт
  (см. dspace generate-sitemaps –h)
- Их никто не прописывает в robots.txt
- С кастомными сайтмапами практически никто не эксперементирует
  (см. http://ideafix.name/?p=1245)

Файлы Sitemap

ДОБАВЛЕНИЕ/ПРОВЕРКА ФАЙЛА SITEMAP

Мои (5) | Все (5)

Содержание файлов Sitemap

Все типы содержания	Веб-страницы
Отправлено	138 908 Отправлено
Проиндексированные	8 558 Проиндексированные

Файлы Sitemap (Все типы содержания)

Загрузить все | Отправить еще раз | Удалить

Показать Строк: 25 1–5 из 5

#	Sitemap	Тип	Обработан	Проблемы	Элементы	Отправлено	Проиндексированные
1	/coolsm.xml	Sitemap	8 нояб. 2013 г.	-	Интернет	4 689	3 977
2	/sitemaps/sitemap_index.xml.gz	Файл индекса Sitemap	7 нояб. 2013 г.	-	Интернет	1 212	1 211
3	/smbig/New sitemap_sitemap.part01.xml.gz	Sitemap	7 нояб. 2013 г.	-	Интернет	50 000	1 688
4	/smbig/New sitemap_sitemap.part02.xml.gz	Sitemap	2 нояб. 2013 г.	-	Интернет	50 000	1 114
5	/smbig/New sitemap_sitemap.part03.xml.gz	Sitemap	7 нояб. 2013 г.	-	Интернет	33 007	568

1–5 из 5

Веселая картинка

WEBMASTER TOOLS

- Google webmaster tools
- Яndex Вебмастер
- Bing веб-мастер
- Baidu?

Google webmaster tools

- Сканирование –> Ошибки сканирования
  Ошибки сервера (5xx)
  Ошибки контента (4xx)
  Ошибки DNS (epic fail)
- Индекс Google –> Статус индексирования
- Проблемы безопасности

- см. http://hdl.handle.net/10995/20792

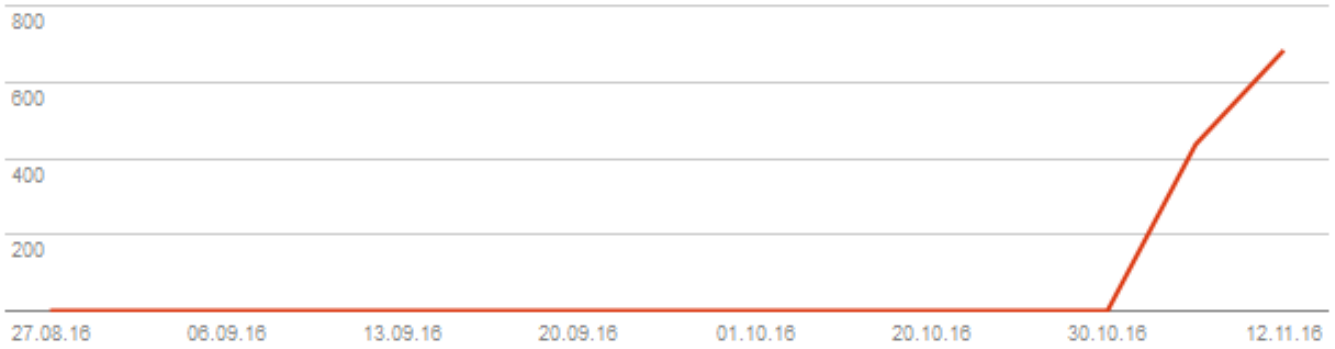
Веселая картинка

Заблокированные ресурсы

Чтобы улучшить индексирование, желательно разрешить обработку некоторых ресурсов.
Подробнее...

Статус на 12.11.16

683 Страницы с заблокированными ресурсами



Содержит | Фильтр

	Хост	Связанные страницы ▼
1	http://elar.uspu.ru	683

Скачать

Показать Строк: 10

1–1 из 1

Продолжение…

Ошибки сайта

Показаны данные за последние 90 дней

DNS | Подключение к серверу | Доступ к файлу robots.txt

Google не удалось получить доступ к сайту из-за ошибки DNS. Подробнее…

- Недоступно
- Ошибка запроса
- Время ожидания истекло
- Общее количество ошибок DNS

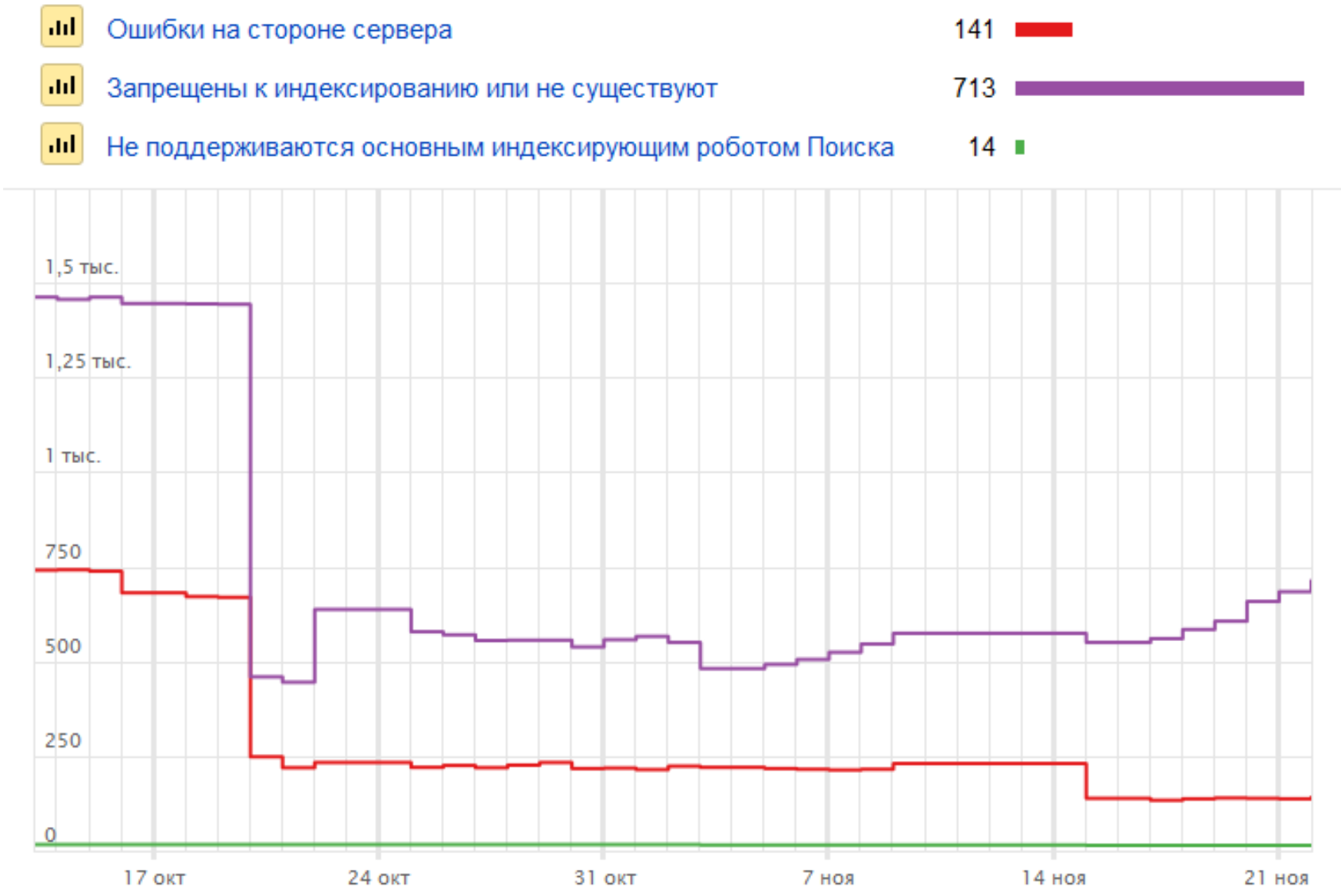
Почему так вышло?

Сетевая служба ВУЗа на неделю положила DNS сервер, перенаправив все запросы на заглушку, которая имела правило Disallow: * в robots.txt

Яndex Вебмастер

- Диагностика
- Индексирование
  статистика
- Настройки индексирования
  Этот момент особенно важен. В нашей практике поисковый бот yandex spider генерирует примерно в 10 раз больше трафика чем google bot, так что, регулярность и интенсивность обхода можно регулировать как в robots.txt, так и через интерфейс вебмастера

Картинка


Ошибки на стороне сервера
141
Запрещены к индексированию или не существуют
713
Не поддерживаются основным индексирующим роботом Поиска
14
1,5 тыс.
1,25 тыс.
1 тыс.
750
500
250
0
17 окт
24 окт
31 окт
7 ноя
14 ноя
21 ноя

Bing веб-мастер

Поисковая система Bing/Yahoo не является популярной у нашей целевой аудитории.

Baidu search

Поисковый робот baidu посещает отечественные репозитории регулярно, но отраженность в поиске от этого не улучшается.

Зарегистрироваться в инструментах вебмастера мне не удалось.

Анализ ошибок

Типы ссылок DSPACE

- **Ссылка на HANDLE – описание:**
  http://aa.aa/handle/P/bb/

- **Ссылка на BITSTREAM – файл:**
  http://aa.aa/bitstream/P/bb/N/X.Y

Ссылка HANDLE

http://aa.aa/handle/P/bb/

1 - префикс протокола

2 - доменное имя

3 - указание на handle

4 - значение префикса

5 - конкретный handle

Ссылка BITSTREAM

http://aa.aa/bitstream/P/bb/N/X.Y

1 - префикс протокола

2 - доменное имя

3 - указание на handle

4 - значение префикса

5 - конкретный handle

6 - номер bitstream

7 - имя документа

Источники ошибок

http://aa.aa/bitstream/P/bb/N/X.Y

1 - смена протокола (на https)

2 - Смена доменного имени

3 - указание на handle/bitstream

4 - Смена префикса

5 - Смена handle (не желательно)

6 - номер bitstream

7 - имя документа (не желательно)

Источники ошибок

- Смена доменного имени
- Смена handle prefix
- Перенос фонда (смена сервера, восстановление бэкапа)
- Изменение версии платформы
- Обновление содердимого (массовая замена PDF файлов)

«Уровни» актуализации ссылок

- Чтобы всё открывалось у нас
- Чтобы всё открывалось еще и из поисковых систем
- Чтобы всё открывалось у всех

Чтобы всё открывалось у нас

- Актуальные значения параметров **handle.canonical.prefix** и **handle.prefix**

- Верные редиректы на уровне WEB-сервера

Чтобы открывалось и из поисковых систем

- Актуальные значения в robots.txt
- Актуальный sitemap
- Ручное удаление неактуальных ссылок
- Ручная актуализация правил robots.txt

Чтобы открывалось отовсюду

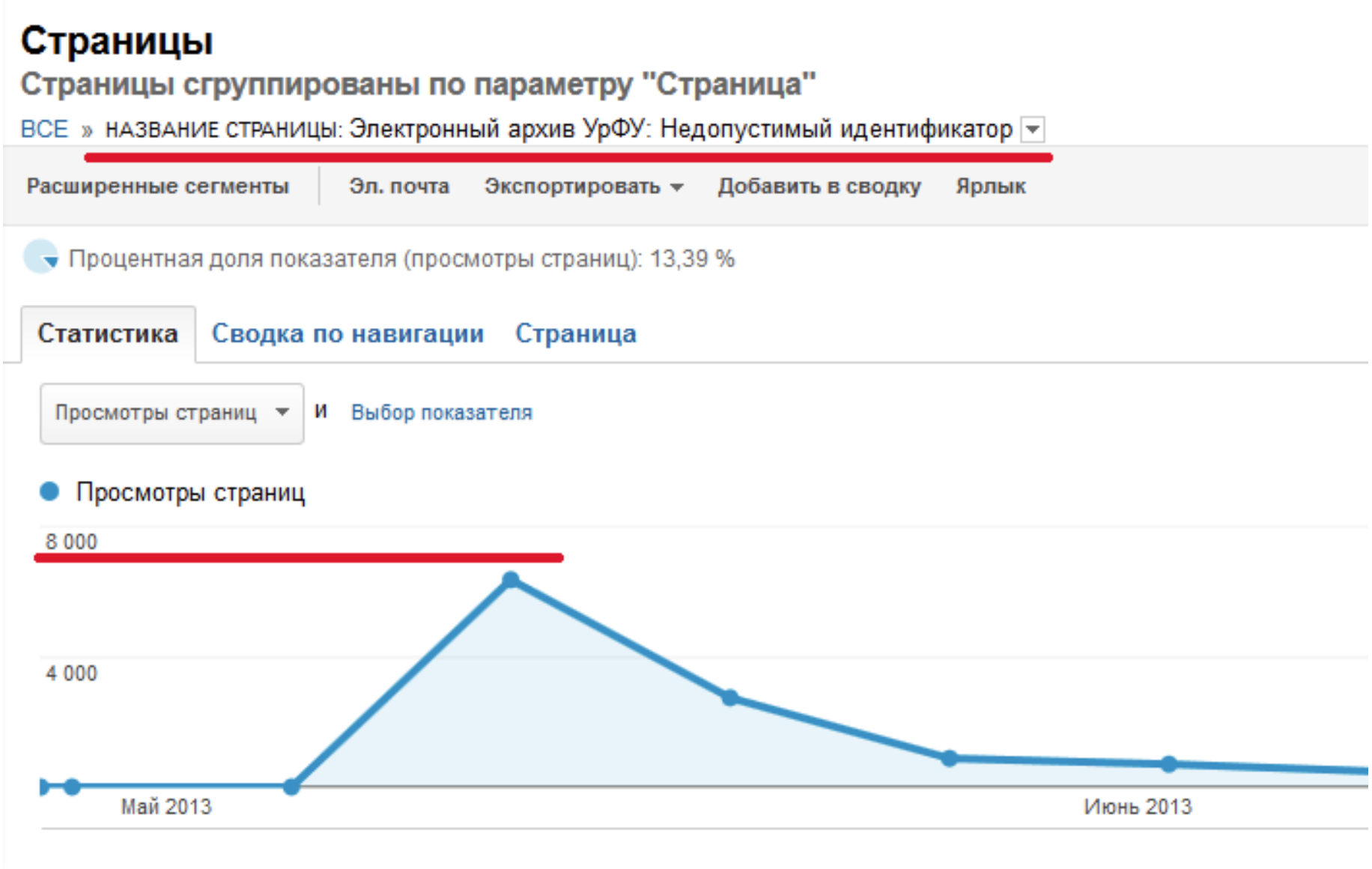
- Ручная актуализация правил редиректов WEB-сервера
- Поиск неактуальных входящих ссылок и их актуализация

Google Analytics, Яndex Метрика
Google webmaster tools, Яndex Вебмастер

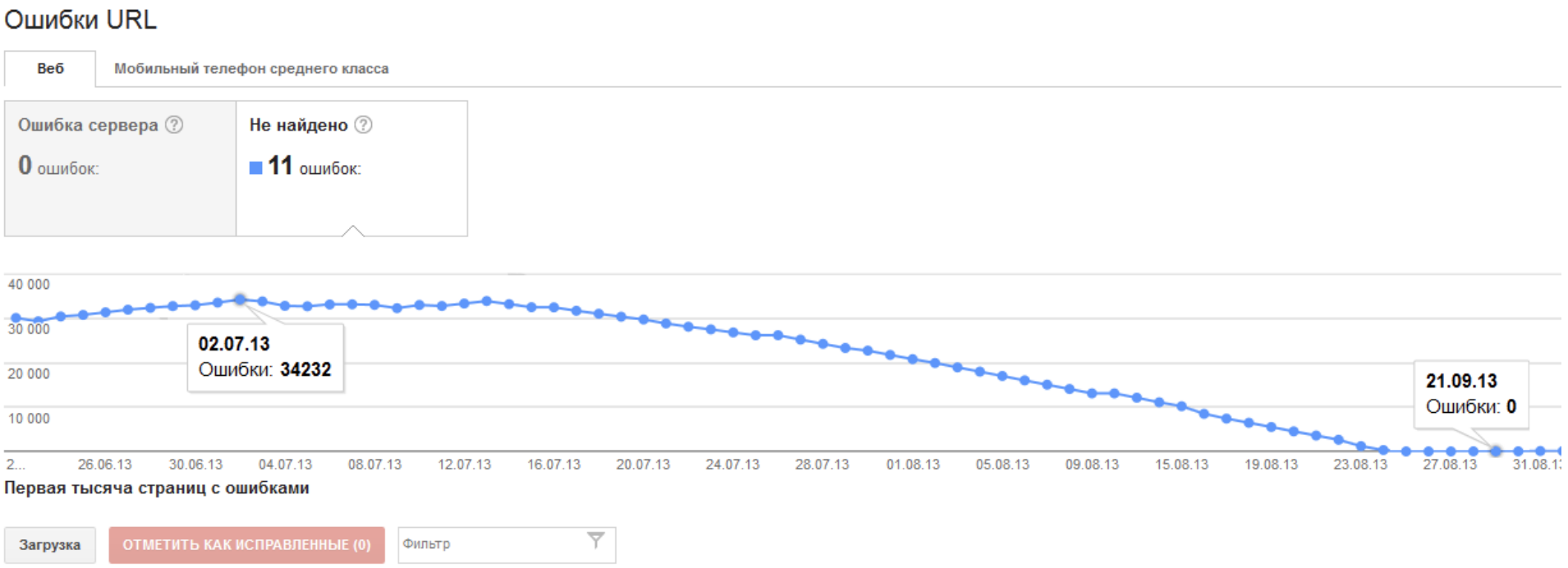
/bitstream/10995/1478/7/1324634_tasks.pdf	Электронный архив УрФУ: Недопустимый идентификатор
/bitstream/10995/1516/4/1331980_exam.pdf	Электронный архив УрФУ: Недопустимый идентификатор
/bitstream/10995/1601/5/1334887_schoolbook.pdf	Электронный архив УрФУ: Недопустимый идентификатор
/bitstream/10995/1724/8/1334956	Электронный архив УрФУ: Недопустимый идентификатор
/bitstream/10995/3963/1/conf_4-5.05.12.pdf	Электронный архив УрФУ: Недопустимый идентификатор

Это плохо, надо принимать меры

30x редиректы



30x редиректы



BONUS – статические «заглушки»

```
<HTML>
<HEAD>
<META HTTP-EQUIV="Content-Type" CONTENT="text/html; charset=UTF-8">
<META HTTP-EQUIV="Refresh" content="7; url=http://elar.usfeu.ru">
<TITLE>Внимание!</TITLE>
</HEAD><BODY>
<center>
<h1>Внимание!</h1>
<p>Ресурс более недоступен. Вы будете перенаправлены на главную страницу архива через 10 секунд.</p>
<p>Если Ваш браузер не поддерживает автоматические переходы, воспользуйтесь <a href="http://elar.usfeu.ru">ссылкой</a></p>
</center>
</BODY>
</HTML>
```