Финансовый университет при Правительстве Российской Федерации

Колледж информатики и программирования

ОТЧЁТ

по лабораторной работе № 8

по МДК.09.02 «Оптимизация веб-приложений»

специальность 09.02.07 «Информационные системы и программирование»

Тема: «Файл robots.txt и карта сайта sitemap.xml»

Проект: интернет-магазин женской одежды Maison Élise

Выполнила:

студентка группы 4ИСИП723

Павлова Дарья Кирилловна

Проверил:

преподаватель Абзалимов Р. Р.

Москва 2026

Цель работы

Настроить управление индексацией сайта Maison Élise: robots.txt, XML-карта сайта с автоматической генерацией, мета-тег robots, канонические адреса и Open Graph.

Задачи

1) определить индексируемые и служебные страницы;

2) составить robots.txt;

3) реализовать генератор sitemap.xml из БД;

4) добавить meta robots, canonical, og-теги;

5) проверить файлы валидатором и собственным скриптом.

Ход выполнения работы

Исходно robots.txt и sitemap.xml отсутствовали (404), не было canonical и meta robots. Индексируемыми определены главная, каталог, акции, «О бренде», «Доставка» и карточки активных товаров; служебные — админка, API, кабинет, корзина, вход, оформление, служебные каталоги.

robots.txt (рисунок 1) сохранён в UTF-8 без BOM, содержит Allow/Disallow, правила для параметров (partial, utm_, ym_internal), Clean-param для UTM, Crawl-delay для агрессивных ботов и строку Sitemap, которую обновляет генератор. В ЛР № 11 правила для личных страниц были пересмотрены: робот должен видеть их noindex.



```
# robots.txt — Maison Élise, интернет-магазин женской одежды
# МДК.09.02, ЛР № 8. Кодировка UTF-8 без BOM.

User-agent: *
# --- Публичные разделы ---
Allow: /$
Allow: /index.php
Allow: /catalog.php
Allow: /product.php
Allow: /promotions.php
Allow: /about.php
Allow: /delivery.php
Allow: /sitemap.xml
Allow: /css/
Allow: /js/
Allow: /images/
# --- Служебные разделы и персональные страницы ---
Disallow: /admin/
Disallow: /api/
Disallow: /app/
Disallow: /sql/
Disallow: /tools/
Disallow: /tests/
Disallow: /integrity/
Disallow: /docs/
Disallow: /images/original/
Disallow: /login.php
Disallow: /register.php
Disallow: /logout.php
Disallow: /cart.php
Disallow: /checkout.php
Disallow: /account.php
Disallow: /favorites.php
Disallow: /generate_sitemap.php
Disallow: /config.php
# --- Дубли: технические параметры ---
Disallow: /*?*partial=
Disallow: /*?*utm_
Disallow: /*?*ym_internal=
Clean-param: utm_source&utm_medium&utm_campaign&utm_content&utm_term /

# Ограничение частоты обхода для агрессивных краулеров
User-agent: AhrefsBot
Crawl-delay: 10

User-agent: SemrushBot
Crawl-delay: 10

Sitemap: http://127.0.0.1:8000/sitemap.xml
```

Рисунок 1 — Файл robots.txt

generate_sitemap.php (листинг 1) формирует карту через XMLWriter: 5 страниц и 6 активных товаров, lastmod по дате изменения файла или товара, priority и changefreq; доступ — только администратор (гость получает 403) или запуск из CLI/cron.

Листинг 1 — Генерация записей карты сайта (generate_sitemap.php)

```
$urls = [
    ['loc' => site_url(''), 'lastmod' => $fileDate('index.php'), 'changefreq'
=> 'daily', 'priority' => '1.0'],
    ['loc' => site_url('catalog.php'), 'lastmod' => $fileDate('catalog.php'),
'changefreq' => 'daily', 'priority' => '0.9'],
    // Задание 6: раздел акций (ЛР № 5) обновляется ежедневно и ведёт на
коммерческие страницы
    ['loc' => site_url('promotions.php'), 'lastmod' => date('Y-m-d'),
'changefreq' => 'daily', 'priority' => '0.8'],
    ['loc' => site_url('about.php'), 'lastmod' => $fileDate('about.php'),
'changefreq' => 'monthly', 'priority' => '0.5'],
    ['loc' => site_url('delivery.php'), 'lastmod' =>
$fileDate('delivery.php'), 'changefreq' => 'monthly', 'priority' => '0.5'],
];
```

В app/seo.php добавлены noindex_pages(), canonical_url() (для товара сохраняется только id, главная — корень) и meta_robots(); шапка выводит meta robots, canonical и og:*. Несуществующий товар отдаёт 404 с noindex.

Найдена и исправлена ошибка: lastmod выводился по UTC (06.11 вместо 07.11) — задан часовой пояс Europe/Moscow.

Результаты (рисунки 2–6).

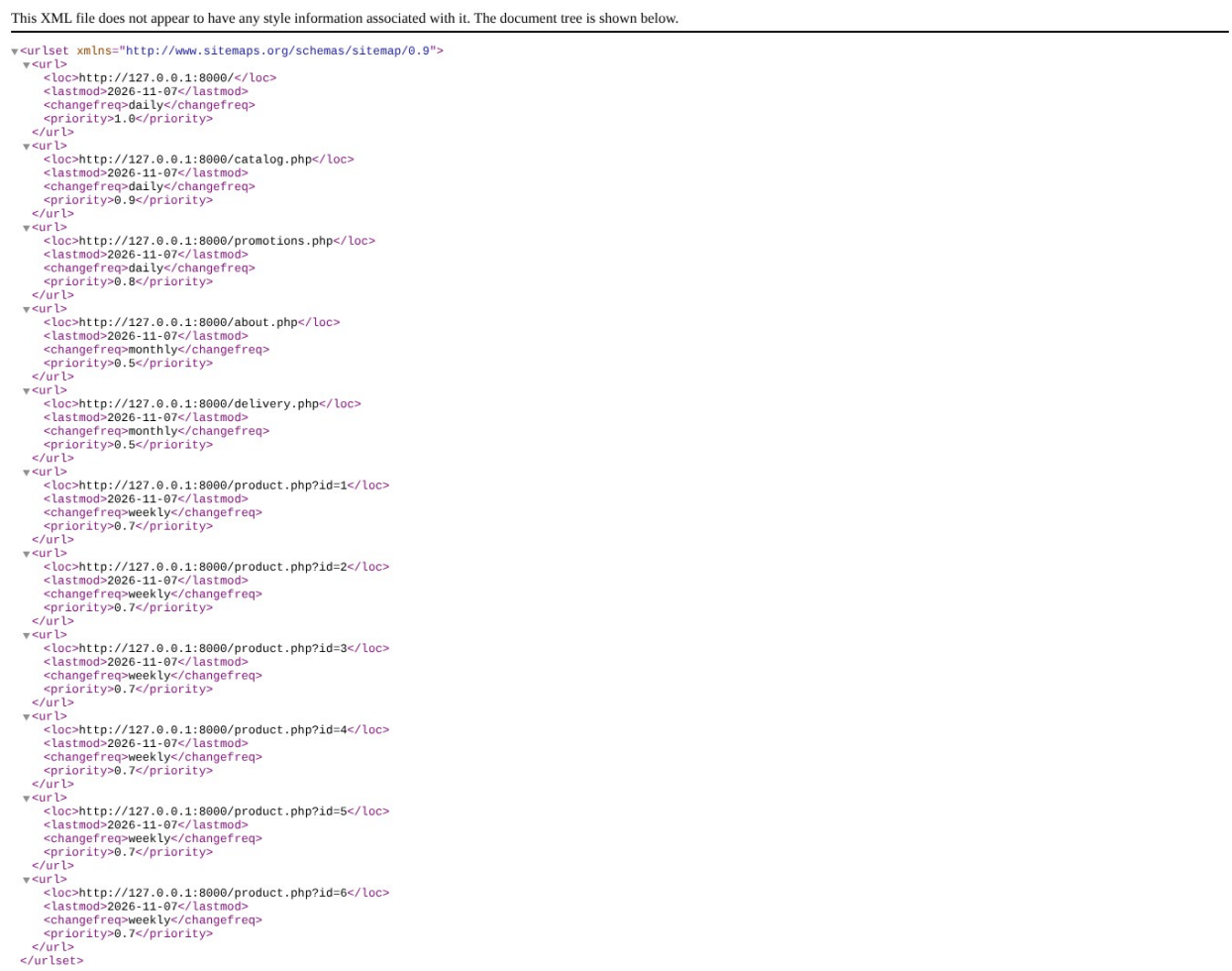


Рисунок 2 — Сгенерированный sitemap.xml

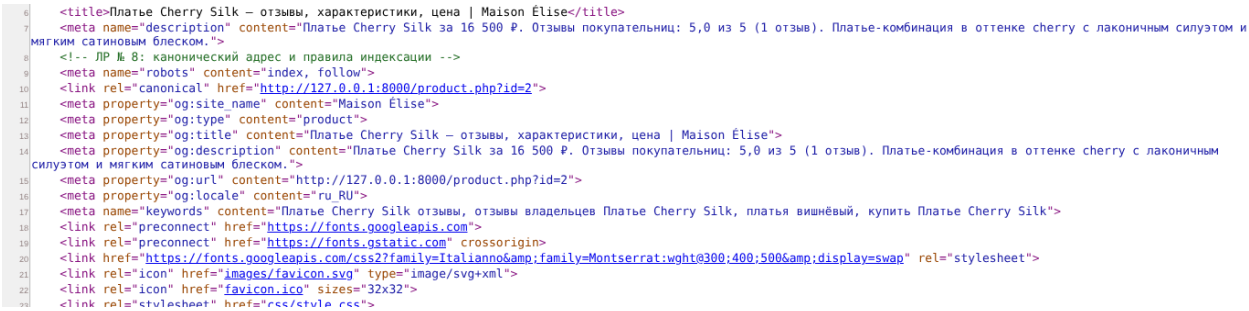


Рисунок 3 — Canonical и meta robots карточки товара

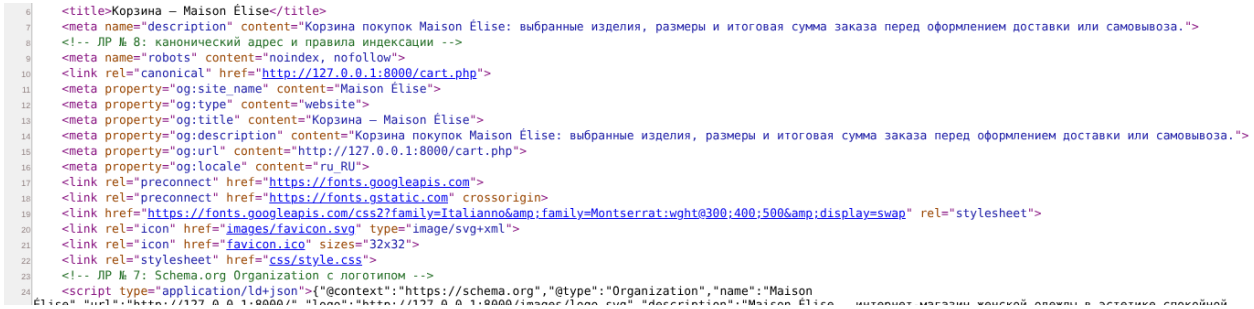


Рисунок 4 — Корзина: noindex, nofollow

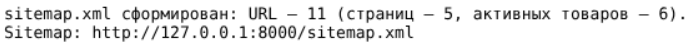


Рисунок 5 — Генерация карты администратором

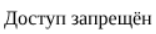


Рисунок 6 — Гость получает 403

Проверка результатов

Карта проверена xmllint по официальной схеме sitemaps.org 0.9 — «sitemap.xml validates» (рисунок 7); 6 URL товаров = 6 активных товаров; Content-Type robots.txt — text/plain, sitemap — application/xml. Скрипт tools/robots_check.py применяет правило самого длинного совпадения и показывает для URL решение robots, meta robots, X-Robots-Tag и canonical (рисунок 8): адрес с utm разрешён (Allow /product.php длиннее), но склеивается canonical и Clean-param.

Вывод команд: генерация и проверка sitemap.xml и robots.txt

```
$ SITE_URL=http://127.0.0.1:8000 php generate_sitemap.php
sitemap.xml сформирован: URL — 11 (страниц — 5, активных товаров — 6).
Sitemap: http://127.0.0.1:8000/sitemap.xml

$ xmllint --noout --schema sitemap.xsd sitemap.xml
sitemap.xml validates

$ grep -c "<url>" sitemap.xml
11

mysql> SELECT COUNT(*) AS active_products FROM products WHERE is_active = 1;
+-----------------+
| active_products |
+-----------------+
|               6 |
+-----------------+
$ grep -c "product.php?id=" sitemap.xml
6

$ curl -sI http://127.0.0.1:8000/robots.txt | grep -i content-type
Content-Type: text/plain; charset=UTF-8
$ curl -sI http://127.0.0.1:8000/sitemap.xml | grep -i content-type
Content-Type: application/xml
$ head -c 3 robots.txt | od -An -tx1   # проверка отсутствия BOM (EF BB BF)
 23 20 72
```

Рисунок 7 — Проверка sitemap.xml по XSD-схеме

Вывод команды: проверка правил robots.txt, meta robots и canonical

```
$ python tools/robots_check.py http://127.0.0.1:8000
robots.txt: правил для User-agent: * — 32
URL                                robots.txt                 HTTP  meta robots         X-Robots-Tag         canonical
/                                  Allow /$                   200   index, follow       —                    /
/index.php                         Allow /index.php           200   index, follow       —                    /
/catalog.php                       Allow /catalog.php         200   index, follow       —                    /catalog.php
/product.php?id=1                  Allow /product.php         200   index, follow       —                    /product.php?id=1
/product.php?id=2&utm_source=vk    Allow /product.php         200   index, follow       —                    /product.php?id=2
/promotions.php                    Allow /promotions.php      200   index, follow       —                    /promotions.php
/about.php                         Allow /about.php           200   index, follow       —                    /about.php
/delivery.php                      Allow /delivery.php        200   index, follow       —                    /delivery.php
/sitemap.xml                       Allow /sitemap.xml         200   —                   —                    —
/catalog_io.php                    Allow (нет правила)        200   noindex, follow     —                    /catalog.php
/cart.php                          Disallow /cart.php         200   noindex, nofollow   —                    /cart.php
/checkout.php                      Disallow /checkout.php     200   noindex, nofollow   —                    /checkout.php
/login.php                         Disallow /login.php        200   noindex, nofollow   —                    /login.php
/register.php                      Disallow /register.php     200   noindex, nofollow   —                    /register.php
/account.php                       Disallow /account.php      200   noindex, nofollow   —                    /login.php
/favorites.php                     Disallow /favorites.php    200   noindex, nofollow   —                    /favorites.php
/admin/index.php                   Disallow /admin/           200   noindex, nofollow   —                    /login.php
/api/products.php                  Disallow /api/             200   —                   noindex              —
/product.php?id=1&partial=1        Allow /product.php         200   —                   noindex              —
/generate_sitemap.php              Disallow /generate_sitemap.p 403   —                   noindex, nofollow    —
```

Рисунок 8 — Проверка правил robots.txt для набора URL

Таблица 1 — Результаты выполнения заданий

№	Задание / результат	Статус	Подтверждение
1	robots.txt	Выполнено и проверено	рис. 1, 8
2	Генератор sitemap.xml	Выполнено и проверено	листинг 1, рис. 2, 5–6
3	Валидация sitemap по XSD	Выполнено и проверено	рис. 7
4	meta robots, canonical, OG	Выполнено и проверено	рис. 3–4
5	Проверка в Яндекс.Вебмастере, xml-sitemaps, оператор site:	Невозможно из-за отсутствия внешнего доступа	сайт не опубликован

Ответы на контрольные вопросы

1. Роль robots.txt и где он находится?

Задаёт правила обхода для роботов; располагается строго в корне сайта (/robots.txt).

2. Disallow и noindex — отличия?

Disallow запрещает сканирование, noindex — индексацию. При Disallow робот не увидит noindex, поэтому страница может остаться в индексе по внешним ссылкам.

3. Директивы частоты обхода и главного зеркала?

Crawl-delay (Яндекс ныне игнорирует, используется настройка в Вебмастере), Host (устарела, заменена 301-редиректом); Clean-param — для параметров.

4. Элементы sitemap?

Обязательны urlset и url/loc; необязательны lastmod, changefreq, priority.

5. changefreq и priority влияют на позиции?

Нет, это подсказки для обхода; на ранжирование не влияют.

6. Как работает приоритет Allow/Disallow?

Побеждает правило с более длинным совпадающим путём; при равенстве — Allow.

7. Почему закрывать служебные страницы?

Они не нужны в выдаче, тратят краулинговый бюджет; дубли размывают релевантность и мешают выбрать правильную страницу.

8. Как помогает canonical?

Указывает основной адрес среди дублей (UTM, сортировки), сигналы склеиваются на нём.

9. Сервисы проверки?

Яндекс.Вебмастер (анализ robots.txt, sitemap), Google Search Console, валидаторы XML-схемы; в работе — xmllint и собственный скрипт.

10. Как правила влияют на скорость индексации?

Корректный sitemap с lastmod ускоряет обнаружение новых страниц, а лишние запреты могут её остановить.

Вывод

Индексация сайта управляется через robots.txt, автоматически генерируемый sitemap.xml, meta robots и canonical. Карта валидна по официальной схеме, правила проверены скриптом. Проверка в сервисах поисковых систем невозможна до публикации сайта.