Финансовый университет при Правительстве Российской Федерации

Колледж информатики и программирования

ОТЧЁТ

по лабораторной работе № 11

по МДК.09.02 «Оптимизация веб-приложений»

специальность 09.02.07 «Информационные системы и программирование»

Тема: «Аудит индексации: разбор исключённых страниц и SQL-журнал состояния индекса»

Проект: интернет-магазин женской одежды Maison Élise

Выполнила:

студентка группы 4ИСИП723

Павлова Дарья Кирилловна

Проверил:

преподаватель Абзалимов Р. Р.

Москва 2026

Цель работы

Провести аудит индексируемости сайта Maison Élise, найти исключённые и дублирующиеся адреса, исправить ошибки и вести SQL-журнал состояния индекса.

Ограничения

Сайт не опубликован, доступа к Яндекс.Вебмастеру нет: подтверждение прав (DNS TXT), «Проверка страницы», переобход и мониторинг Вебмастера не выполнялись. Вместо данных Вебмастера используется локальный обходчик; его записи помечены source = 'local_crawl' и не выдаются за данные Яндекса.

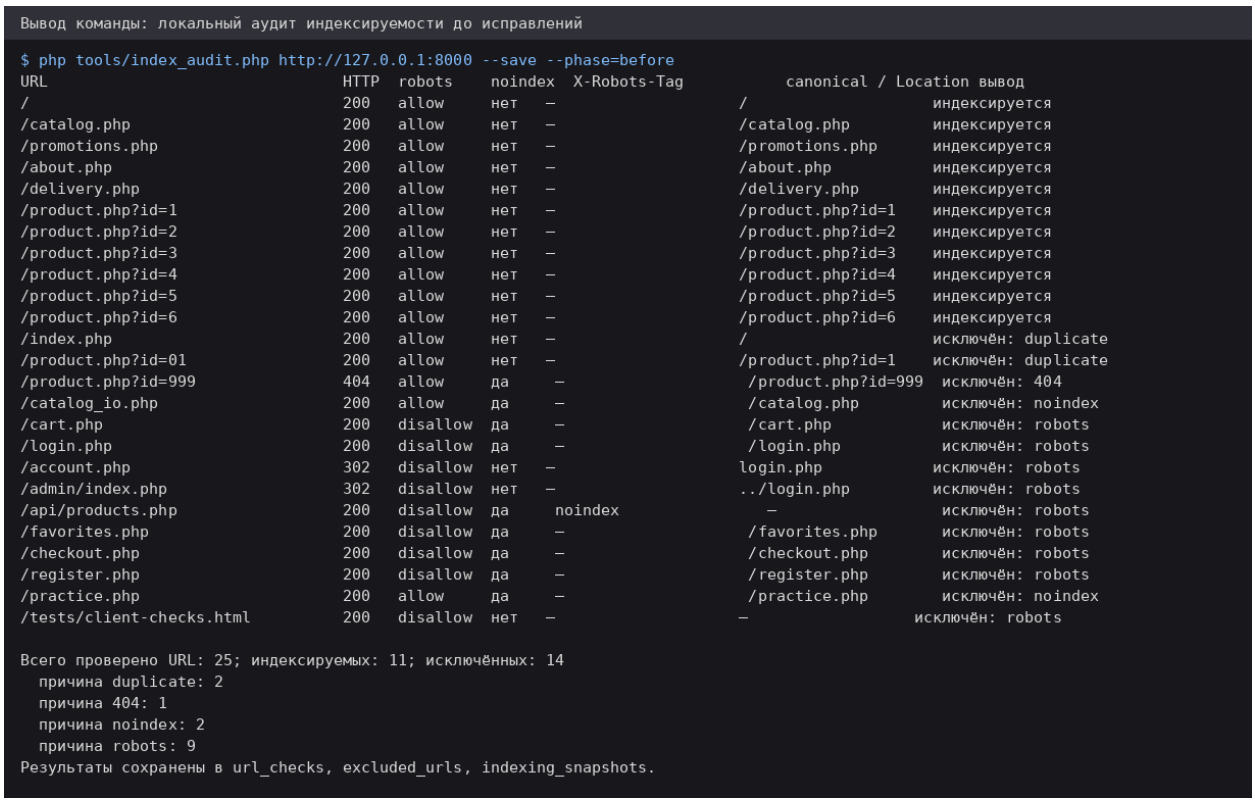
Ход выполнения работы

Миграция 006_lr11_indexing.sql создала таблицы indexing_snapshots (источник yandex/google/local_crawl, фаза before/after/monitoring), excluded_urls (причина, действие, статус, признак намеренного исключения) и url_checks. Скрипт tools/index_audit.php обходит сайт с главной и из sitemap, фиксирует код ответа, meta robots и X-Robots-Tag, canonical, решение robots.txt и причину исключения (листинг 1).

Листинг 1 — Классификация причины исключения (tools/index_audit.php)

```
    $robotsVerdict = robots_decide($path, $rules);
    $reason = null;
    if ($robotsVerdict === 'disallow') $reason = 'robots';
    elseif ($r['code'] >= 300 && $r['code'] < 400) $reason = 'redirect';
    elseif ($r['code'] === 404) $reason = '404';
    elseif ($noindex) $reason = 'noindex';
    elseif ($canonical !== null && $canonical !== $path && $canonical !==
rtrim($path, '/')) $reason = 'duplicate';
    elseif ($r['code'] === 200 && str_contains($r['body'], 'Товар не
найден')) $reason = 'soft404';
```

До исправлений: 25 URL, 11 индексируемых, 14 исключённых — robots 9 (личные страницы и служебные разделы), дубли 2 (/index.php и product.php?id=01), 404 — 1, noindex — 2 (рисунок 1). Проблема: личные страницы были закрыты Disallow, поэтому робот не видел их noindex.



```
Вывод команды: локальный аудит индексируемости до исправлений

$ php tools/index_audit.php http://127.0.0.1:8000 --save --phase=before
URL                              HTTP  robots    noindex  X-Robots-Tag          canonical / Location вывод
/                                200   allow     нет   –                  /                    индексируется
/catalog.php                     200   allow     нет   –                  /catalog.php         индексируется
/promotions.php                  200   allow     нет   –                  /promotions.php      индексируется
/about.php                       200   allow     нет   –                  /about.php           индексируется
/delivery.php                    200   allow     нет   –                  /delivery.php        индексируется
/product.php?id=1                200   allow     нет   –                  /product.php?id=1    индексируется
/product.php?id=2                200   allow     нет   –                  /product.php?id=2    индексируется
/product.php?id=3                200   allow     нет   –                  /product.php?id=3    индексируется
/product.php?id=4                200   allow     нет   –                  /product.php?id=4    индексируется
/product.php?id=5                200   allow     нет   –                  /product.php?id=5    индексируется
/product.php?id=6                200   allow     нет   –                  /product.php?id=6    индексируется
/index.php                       200   allow     нет   –                  /                    исключён: duplicate
/product.php?id=01               200   allow     нет   –                  /product.php?id=1    исключён: duplicate
/product.php?id=999              404   allow     да     –                  /product.php?id=999  исключён: 404
/catalog_io.php                  200   allow     да     –                  /catalog.php         исключён: noindex
/cart.php                        200   disallow  да     –                  /cart.php            исключён: robots
/login.php                       200   disallow  да     –                  /login.php           исключён: robots
/account.php                     302   disallow  нет   –                  login.php            исключён: robots
/admin/index.php                 302   disallow  нет   –                  ../login.php         исключён: robots
/api/products.php                200   disallow  да     noindex              –                  исключён: robots
/favorites.php                   200   disallow  да     –                  /favorites.php       исключён: robots
/checkout.php                    200   disallow  да     –                  /checkout.php        исключён: robots
/register.php                    200   disallow  да     –                  /register.php        исключён: robots
/practice.php                    200   allow     да     –                  /practice.php        исключён: noindex
/tests/client-checks.html        200   disallow  нет   –                  –                  исключён: robots

Всего проверено URL: 25; индексируемых: 11; исключённых: 14
  причина duplicate: 2
  причина 404: 1
  причина noindex: 2
  причина robots: 9
Результаты сохранены в url_checks, excluded_urls, indexing_snapshots.
```

Рисунок 1 — Аудит индексации до исправлений

Исправления: из robots.txt удалены Disallow для входа, регистрации, корзины, оформления, кабинета и избранного; noindex дублируется заголовком X-Robots-Tag; функция enforce_canonical_url() делает 301 с /index.php на / и с product.php?id=01 на ?id=1 (листинг 2). Найдена ошибка: первая версия обрабатывала только GET, и curl -I получал 200 — добавлен HEAD.

Листинг 2 — 301-редирект дублей (app/seo.php)

```
function enforce_canonical_url(): void
{
    if (headers_sent() || !in_array($_SERVER['REQUEST_METHOD'] ?? 'GET',
['GET', 'HEAD'], true)) {
        return;
    }
    $uriPath = parse_url($_SERVER['REQUEST_URI'] ?? '/', PHP_URL_PATH) ?:
'/';
    $page = current_page();
    $target = null;
    if (($page === 'index.php' || $page === 'en/index.php') &&
str_ends_with($uriPath, '/index.php')) {
        $query = $_SERVER['QUERY_STRING'] ?? '';
        $target = substr($uriPath, 0, -strlen('index.php')) . ($query !==
'' ? '?' . $query : '');
    }
    if (($page === 'product.php' || $page === 'en/product.php') &&
isset($_GET['id']) && !isset($_GET['partial'])) {
        $raw = (string) $_GET['id'];
        $id = (int) $raw;
```

```
        if ($id > 0 && $raw !== (string) $id) {
            $params = $_GET;
            $params['id'] = $id;
            $target = app_base_path() . '/' . $page . '?' .
http_build_query($params);
        }
    }
    if ($target !== null) {
        header('Location: ' . $target, true, 301);
        exit;
```

Результаты (рисунки 2–5).

Вывод curl -I: коды ответов, 301-редиректы и X-Robots-Tag

```
$ curl -sI http://127.0.0.1:8000/
HTTP/1.1 200 OK
$ curl -sI http://127.0.0.1:8000/index.php
HTTP/1.1 301 Moved Permanently
Location: /
$ curl -sI http://127.0.0.1:8000/index.php?utm_source=vk
HTTP/1.1 301 Moved Permanently
Location: /?utm_source=vk
$ curl -sI http://127.0.0.1:8000/catalog.php
HTTP/1.1 200 OK
$ curl -sI http://127.0.0.1:8000/product.php?id=1
HTTP/1.1 200 OK
$ curl -sI http://127.0.0.1:8000/product.php?id=01
HTTP/1.1 301 Moved Permanently
Location: /product.php?id=1
$ curl -sI http://127.0.0.1:8000/product.php?id=999
HTTP/1.1 404 Not Found
X-Robots-Tag: noindex, follow
$ curl -sI http://127.0.0.1:8000/cart.php
HTTP/1.1 200 OK
X-Robots-Tag: noindex, nofollow
$ curl -sI http://127.0.0.1:8000/login.php
HTTP/1.1 200 OK
X-Robots-Tag: noindex, nofollow
$ curl -sI http://127.0.0.1:8000/account.php
HTTP/1.1 302 Found
Location: login.php
$ curl -sI http://127.0.0.1:8000/admin/index.php
HTTP/1.1 302 Found
Location: ../login.php
$ curl -sI http://127.0.0.1:8000/api/products.php
HTTP/1.1 200 OK
X-Robots-Tag: noindex
$ curl -sI http://127.0.0.1:8000/robots.txt
HTTP/1.1 200 OK
$ curl -sI http://127.0.0.1:8000/sitemap.xml
HTTP/1.1 200 OK
```

Рисунок 2 — Коды ответов и заголовки (curl -I)

Вывод команды: повторный аудит после исправлений

```
$ php tools/index_audit.php http://127.0.0.1:8000 --save --phase=after
URL                              HTTP  robots    noindex  X-Robots-Tag           canonical / Location вывод
/                                200   allow     нет   –                  /                    индексируется
/catalog.php                     200   allow     нет   –                  /catalog.php         индексируется
/promotions.php                  200   allow     нет   –                  /promotions.php      индексируется
/about.php                       200   allow     нет   –                  /about.php           индексируется
/delivery.php                    200   allow     нет   –                  /delivery.php        индексируется
/product.php?id=1                200   allow     нет   –                  /product.php?id=1    индексируется
/product.php?id=2                200   allow     нет   –                  /product.php?id=2    индексируется
/product.php?id=3                200   allow     нет   –                  /product.php?id=3    индексируется
/product.php?id=4                200   allow     нет   –                  /product.php?id=4    индексируется
/product.php?id=5                200   allow     нет   –                  /product.php?id=5    индексируется
/product.php?id=6                200   allow     нет   –                  /product.php?id=6    индексируется
/index.php                       301   allow     нет   –                  /                    исключён: redirect
/product.php?id=01               301   allow     нет   –                  /product.php?id=1    исключён: redirect
/product.php?id=999              404   allow     да     noindex, follow       /product.php?id=999  исключён: 404
/catalog_io.php                  200   allow     да     noindex, follow       /catalog.php         исключён: noindex
/cart.php                        200   allow     да     noindex, nofollow     /cart.php            исключён: noindex
/login.php                       200   allow     да     noindex, nofollow     /login.php           исключён: noindex
/account.php                     302   allow     нет   –                  login.php            исключён: redirect
/admin/index.php                 302   disallow  нет   –                  ../login.php         исключён: robots
/api/products.php                200   disallow  да     noindex               –                    исключён: robots
/favorites.php                   200   allow     да     noindex, nofollow     /favorites.php       исключён: noindex
/checkout.php                    200   allow     да     noindex, nofollow     /checkout.php        исключён: noindex
/register.php                    200   allow     да     noindex, nofollow     /register.php        исключён: noindex
/practice.php                    200   allow     да     noindex, nofollow     /practice.php        исключён: noindex
/tests/client-checks.html        200   disallow  нет   –                  –                    исключён: robots

Всего проверено URL: 25; индексируемых: 11; исключённых: 14
  причина redirect: 3
  причина 404: 1
  причина noindex: 7
  причина robots: 3
Результаты сохранены в url_checks, excluded_urls, indexing_snapshots.
Мониторинг: изменение числа индексируемых URL +0.0 % (порог 5 %) — в норме
```

Рисунок 3 — Аудит после исправлений

Консоль MySQL 8.4: снимки индекса, исключённые URL и проверки страниц

```
mysql> source sql/lr11_indexing_report.sql
+-------------+--------+----------------+-----------------+--------+---------------+----------------+-------+
| before_date | p1     | indexed_before | excluded_before | p2     | indexed_after | excluded_after | delta |
+-------------+--------+----------------+-----------------+--------+---------------+----------------+-------+
| 2026-11-07  | before |             11 |              14 | after  |            11 |             14 |     0 |
+-------------+--------+----------------+-----------------+--------+---------------+----------------+-------+
+-----------+-----+-----------+---------+-----------------+
| reason    | cnt | fixed_cnt | new_cnt | intentional_cnt |
+-----------+-----+-----------+---------+-----------------+
| robots    |   9 |         6 |       0 |               9 |
| noindex   |   7 |         0 |       0 |               7 |
| redirect  |   3 |         0 |       0 |               3 |
| duplicate |   2 |         2 |       0 |               0 |
| 404       |   1 |         0 |       0 |               1 |
+-----------+-----+-----------+---------+-----------------+
+--------------------+-----------+-------------+--------------------+------------+--------------------+--------------------+---------------------+
| url                | http_code | has_noindex | x_robots_tag       | robots_txt | canonical          | redirect_to        | checked_at          |
+--------------------+-----------+-------------+--------------------+------------+--------------------+--------------------+---------------------+
| /                  |       200 |           0 | NULL               | allow      | /                  | NULL               | 2026-11-07 11:05:27 |
| /cart.php          |       200 |           1 | noindex, nofollow  | allow      | /cart.php          | NULL               | 2026-11-07 11:05:27 |
| /catalog.php       |       200 |           0 | NULL               | allow      | /catalog.php       | NULL               | 2026-11-07 11:05:27 |
| /index.php         |       301 |           0 | NULL               | allow      | NULL               | /                  | 2026-11-07 11:05:27 |
| /product.php?id=01 |       301 |           0 | NULL               | allow      | NULL               | /product.php?id=1  | 2026-11-07 11:05:27 |
| /product.php?id=1  |       200 |           0 | NULL               | allow      | /product.php?id=1  | NULL               | 2026-11-07 11:05:27 |
+--------------------+-----------+-------------+--------------------+------------+--------------------+--------------------+---------------------+
```

Рисунок 4 — SQL-отчёт: снимки и причины исключения

```
# 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: /marketing/
Disallow: /generate_sitemap.php
Disallow: /config.php
# Личные страницы (вход, корзина, кабинет, избранное) НЕ закрыты здесь намеренно (ЛР № 11):
# робот должен увидеть на них meta robots и заголовок X-Robots-Tag «noindex, nofollow».
# --- 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 5 — robots.txt после исправления

Таблица 1 — Состояние до и после (локальный обход)

Показатель	**До**	**После**
Всего URL	25	25
Индексируемые	11	11
Исключены robots	9	3 (админка, API, тесты — намеренно)
Исключены noindex	2	7
Редиректы 301	0	3
Дубли	2	0
404	1	1 (id=999 — намеренно)

Таблица 2 — Результаты выполнения заданий

№	**Задание / результат**	**Статус**	**Подтверждение**
1	SQL-журнал индекса	Выполнено и проверено	рис. 4
2	Локальный аудит исключённых URL	Выполнено и проверено	рис. 1, 3
3	Исправление robots/noindex/дублей	Выполнено и проверено	рис. 2, 5, табл. 1
4	Вебмастер: права, проверка страниц, переобход, мониторинг	Невозможно из-за отсутствия внешнего доступа	нет доступа, сайт не опубликован

Ответы на контрольные вопросы

1. Обход и индексация?

Обход — загрузка страницы роботом, индексация — включение в базу поиска. Страница 200 может не попасть в выдачу из-за noindex, дубля, низкого качества.

2. Disallow и noindex — почему не комбинируют?

Disallow запрещает загрузку, поэтому робот не прочитает noindex на странице.

3. Почему Disallow не удаляет страницу?

Адрес может остаться в индексе по внешним ссылкам; для удаления нужен noindex (страница открыта для обхода) или 404/410.

4. Когда X-Robots-Tag удобнее?

Для не-HTML файлов (PDF, изображения), ответов API и единообразной настройки на уровне сервера.

5. Причины исключения и действия?

Дубль — canonical/301, noindex — проверить намеренность, 404 — исправить ссылки или оставить намеренно, редирект — нормален, запрет robots — снять для нужных страниц, малоценная — улучшить контент.

6. «Проверка страницы» и что логировать?

Код ответа, дату обхода, статус индексации, canonical, meta robots — в url_checks.

7. Зачем indexing_snapshots?

Хранит количество индексируемых URL по датам и фазам — видна динамика и резкие падения (порог 5 %).

8. Связь Вебмастера и Метрики?

Переходы из поиска в Вебмастере сопоставляются с визитами из поисковых систем в Метрике по тем же страницам.

9. Почему нельзя массово переиндексировать?

Есть суточные лимиты, массовые заявки не ускоряют индексацию и тратят квоту.

10. Как отличить намеренное исключение?

Поле intentional и reason в excluded_urls и данные url_checks (noindex на служебной странице — норма, на товаре — ошибка).

Вывод

Локальный аудит выявил ошибочную комбинацию Disallow и noindex и два дубля; после исправлений дублей нет, личные страницы корректно исключены noindex и X-Robots-Tag, служебные — robots.txt. SQL-журнал позволяет отслеживать динамику. Работа с Вебмастером возможна после публикации сайта.