



Факультет
Компьютерных наук

Программная инженерия

Разработка веб-приложения для комплексной автоматизации процессов поискового продвижения веб-сайтов

Audit Service, Semantic Service, Reporting
Service, Асинхронная событийная модель,
Data/AI-инфраструктура

Командный программный проект

Выполнила студентка 2 курса Пухова Александра Игоревна БПИ245

Научный руководитель:
Доцент: Факультет компьютерный
наук/Департамент программной инженерии
Нестеров Роман Александрович

Сокомандник: Зобов Арсений Александрович БПИ245



Описание предметной области

ПОИСКОВАЯ ОПТИМИЗАЦИЯ (SEO)

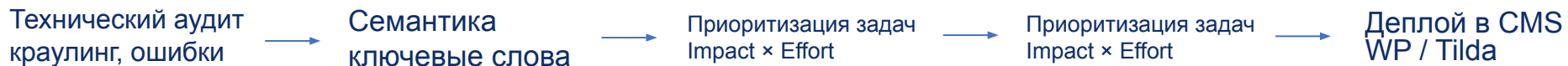
Комплекс мер по повышению позиций веб-сайта в органической выдаче поисковых систем (Google, Яндекс). Включает технический аудит, работу с контентом, ссылочный профиль и мониторинг позиций.

РЫНОК SEO-ИНСТРУМЕНТОВ

Делится на три сегмента: аналитические платформы (Ahrefs, Semrush), технические краулеры (Screaming Frog) и системы автоматизации продвижения (PromoPult, Rookee). Ни один сегмент не охватывает полный цикл.

ЦЕЛЕВАЯ АУДИТОРИЯ

Малый и средний бизнес, владельцы сайтов на WordPress и Tilda, SEO-специалисты — те, кому недоступны Enterprise-решения, но необходима автоматизация полного цикла продвижения.





Основные термины и сокращения

Термин	Расшифровка	Использование в проекте
SEO	Search Engine Optimization	Оптимизация сайта для поискового продвижения
CWV	Core Web Vitals	Качество загрузки и взаимодействия
GSC / GA4	Google Search Console / Analytics	Внешние источники данных
Schema.org	Структурированная разметка	Проверка и генерация JSON-LD
LLM	Large Language Model	Генерация title, description, schema
FF-Score	Freshness/Familiarity/Quality Score	Интегральная метрика SEO-состояния
E-E-A-T	Experience, Expertise, Authoritativeness, Trustworthiness	Оценка качества и доверия контента
HITL	Human-in-the-Loop	Ручное подтверждение критических изменений



Актуальность работы

ПРОБЛЕМА 1 — СТОИМОСТЬ

Коммерческие SEO-инструменты (Ahrefs, Semrush, Similarweb) — от \$99–499/мес. Для малого и среднего бизнеса недоступны.

ПРОБЛЕМА 2 — КОНТРОЛЬ

Инструменты либо полностью ручные (Screaming Frog), либо автоматизированы без контроля критических изменений → риск SEO-катастроф.

ПРОБЛЕМА 3 — РАЗРЫВ КОНТУРА

Нет единого пути от рекомендации до внедрения. Рынок закрывает SEO «по частям»: аудит, задачи и CMS — отдельные продукты.





Цель и задачи

ЦЕЛЬ (ПУХОВА А.И.)

Разработать аналитическое backend-ядро микросервисного веб-приложения, обеспечивающее технический SEO-аудит сайтов, семантический анализ, расчёт FF-Score и E-E-A-T, генерацию LLM-рекомендаций и формирование CSV-отчётности.

ЗАДАЧИ

- Разработка Audit Service: публичный и полный аудит сайта.
- Реализация краулинга Scrapy + Playwright для статических и JS/SPA-страниц.
- Разработка Semantic Service: FF-Score, E-E-A-T, Content Gap.
- Интеграция LLM: OpenAI primary + Gemini fallback.
- Разработка Reporting Service: GSC/GA4/Yandex, pandas, CSV.
- Обеспечение событий CrawlCompleted и FFScoreRecalculated.

ИЗ ЗОНЫ ОТВЕТСТВЕННОСТИ ИСКЛЮЧЕНЫ

- API Gateway и JWT-авторизация.
- Frontend SPA и HITL-UI.
- Management Service и приоритизация задач.
- Client API Gateway, HMAC-deploy и rollback.
- WordPress/Tilda адаптеры внедрения изменений.

ЗОНА ОТВЕТСТВЕННОСТИ ЗОБОВА А.А.

- Management Service
- API Gateway
- Frontend SPA
- Client API Gateway
- WordPress и Tilda
- Обеспечение событийного взаимодействия



Требования к программному продукту

ФУНКЦИОНАЛЬНЫЕ ТРЕБОВАНИЯ

- Публичный аудит по URL без авторизации.
- Полный аудит проекта для зарегистрированного режима.
- Краулинг статических и JS/SPA-страниц.
- Анализ title, description, H1–H6, ссылок, robots.txt, sitemap.xml, Schema.org.
- Получение CWV через PageSpeed Insights.
- Расчёт FF-Score и E-E-A-T Score.
- Генерация LLM-черновиков для title, description, Schema.org.
- Формирование CSV-отчётов.

НЕФУНКЦИОНАЛЬНЫЕ ТРЕБОВАНИЯ

- Асинхронность: Celery/RabbitMQ для долгих задач.
- Масштабируемость: независимые сервисы и воркеры.
- Надёжность: retry, timeout, обработка 4xx/5xx.
- Кэширование: Redis TTL 7 дней для LLM-ответов.
- Хранение: PostgreSQL JSONB для сырых данных аудита.
- Контейнеризация: Docker + Docker Compose.
- Тестируемость: pytest и интеграционные проверки.

ОГРАНИЧЕНИЯ

- Автоматическое внедрение изменений в CMS относится к зоне сокомандника; в индивидуальной части формируются аналитика, метрики, черновики и отчёты.



Анализ существующих решений — аналитика и краулинг

Критерий	Ahrefs	Semrush	MegaIndex	Screaming Frog	Netpeak Spider	Наша разработка
Публичный API данных	+	+	+	-	-	+
Мониторинг позиций	+	+	+	-	+	+
JS/SPA краулинг	-	-	-	+	+	+
Автоматизация запусков	-	Частично	-	+	-	+
Параллельная работа с проектами	+	+	+	-	+	+
НITL-контроль изменений	-	-	-	-	-	+
Безопасное внедрение в CMS	-	-	-	-	-	+



Анализ существующих решений — автоматизация продвижения

Критерий	PromoPult	Rooke	WebEffector	REG.RU Seowizard	Timeweb	Наша разработка
Запуск из ЛК/панели	+	+	+	+	+	+
Аудит + рекомендации	Частично	+	+	+	Частично	+
НITL ручной контроль	-	-	-	-	-	+
Безопасное внедрение в CMS	-	-	-	-	-	+
Мониторинг позиций	+	+	+	+	-	+
Прозрачная аналитика (FF-Score)	Частично	+	+	+	-	+



Функционал индивидуальной части

AUDIT SERVICE

- Публичный и полный SEO-аудит сайта
- Краулинг Scrapy + Playwright
- JS/SPA-рендеринг страниц
- Анализ meta, H1–H6, ссылок
- Проверка robots.txt, sitemap.xml, Schema.org
- CWV через PageSpeed Insights
- CrawlCompleted event

SEMANTIC SERVICE

- Расчёт FF-Score
- Расчёт E-E-A-T Score
- Контент-анализ и Content Gap
- LLM drafts: title, description, Schema.org
- OpenAI primary + Gemini fallback
- Redis cache TTL 7 дней
- FFScoreRecalculated event

REPORTING SERVICE

- GSC / GA4 / Yandex collectors
- pandas-агрегация метрик
- ROI и HITL efficiency
- report_snapshot
- CSV export с динамическими колонками
- PostgreSQL JSONB

Режимы оркестрации по FF-Score:

FF-Score < 40 — «Спасение»
CWV · 404 · robots.txt · Schema.org

$40 \leq FF \leq 60$ — «Баланс»
Технические + контентные задачи

FF-Score > 60 — «Пост»
Content Gap · E-E-A-T · LLM-рекомендации

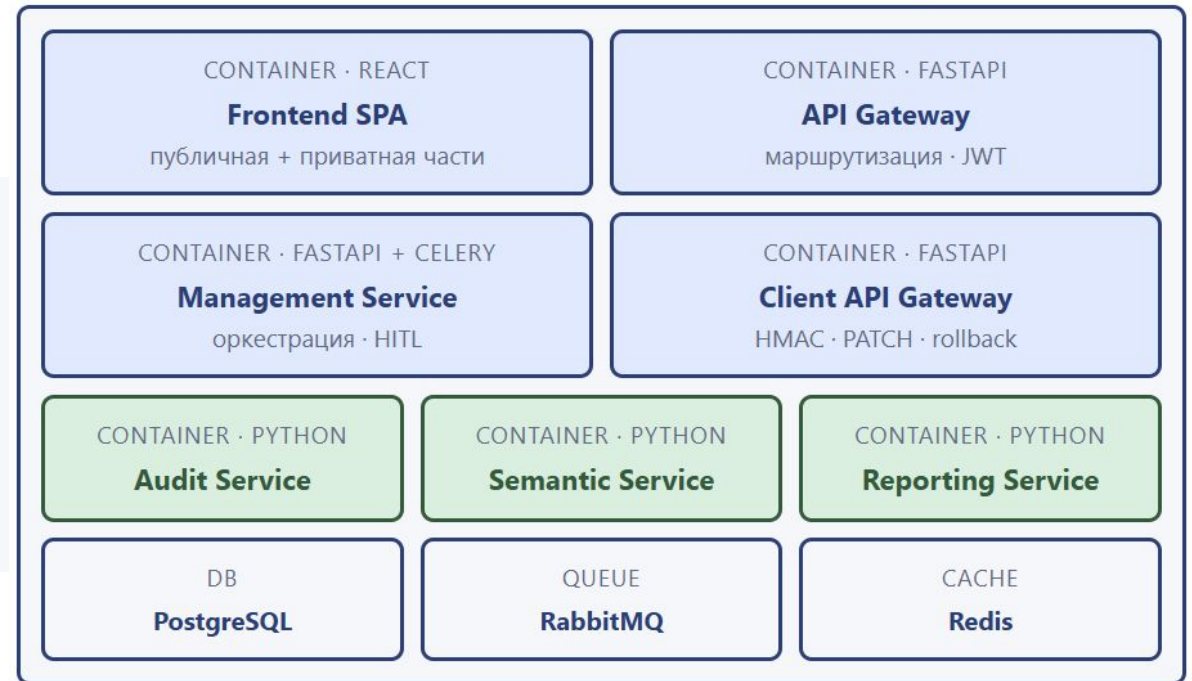


Уровни контекст и контейнера

CONTEXT — СИСТЕМА И ВНЕШНИЕ АКТОРЫ



CONTAINER — КЛЮЧЕВЫЕ КОНТЕЙНЕРЫ (ЗОНА ПУХОВА А.И. ВЫДЕЛЕНА)



□ Зона Зобов А.А. □ Зона Пухова А.И.



Уровни контекст и контейнера

CONTEXT — СИСТЕМА И ВНЕШНИЕ АКТОРЫ



CONTAINER — КЛЮЧЕВЫЕ КОНТЕЙНЕРЫ





Audit Service — алгоритм краулинга и технического аудита

ПУБЛИЧНЫЙ КРАУЛЕР

- Вход: URL сайта без авторизации.
- Ограничения: до 10 страниц, глубина 2.
- Проверки: HTTP-статусы, title, description, H1, ссылки, базовые ошибки.
- Rate limit: 5 запросов/час/IP.
- Результат: public_audit_results, retention 7 дней.

ПОЛНЫЙ КРАУЛЕР

- - Scrapy — быстрый обход статических страниц.
- Playwright — DOM после JS-rendering для SPA.
- Pool Playwright: до 5 инстансов, timeout 30 сек.
- CWV через PageSpeed Insights.
- Schema.org validation.
- Результат: crawl_results JSONB + событие CrawlCompleted.



Semantic Service — FF-Score и LLM-контур

FF-SCORE

$$FF = 0.30 \times \text{Freshness} + 0.40 \times \text{Familiarity} + 0.30 \times \text{Quality}$$

- Диапазон: 0–100.
- Режимы: Rescue / Balanced / Growth.
- Контрольный набор: FF-Score = 69.1.
- Freshness = 52.3, Familiarity = 66.0, Quality = 90.0.

LLM И E-E-A-T

- OpenAI — primary provider.
- Google Gemini — fallback provider.
- temperature = 0.4, timeout = 15 сек.
- Redis cache: TTL 7 дней.
- Prompts: TITLE, DESCRIPTION, SCHEMA.
- Результат: LLM drafts для HITL, без автопубликации.



Reporting Service — отчётность и CSV

ИСТОЧНИКИ И АГРЕГАЦИЯ

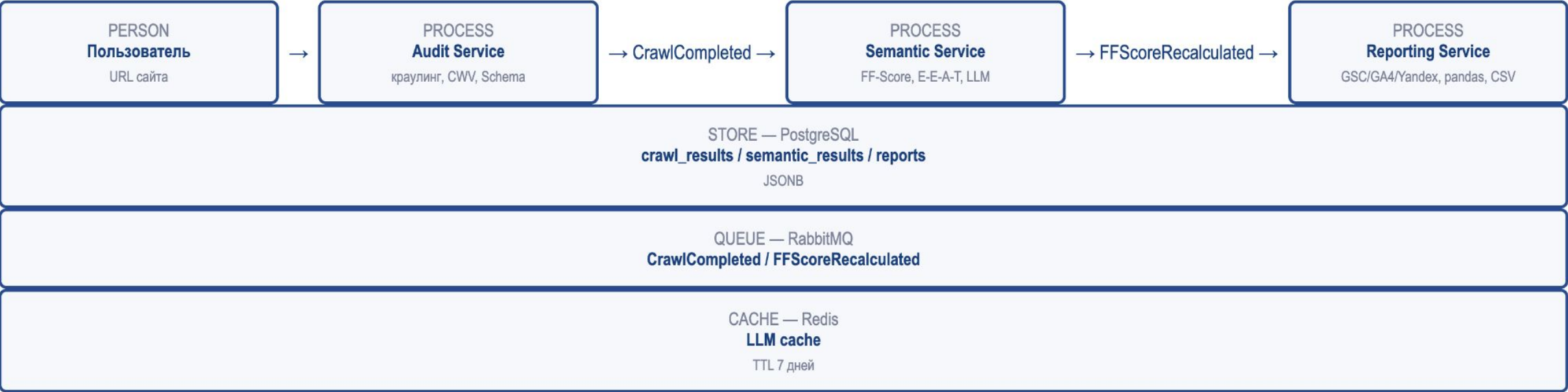
- GSC collector: клики, показы, CTR, позиции.
- GA4 collector: sessions, users, конверсии.
- Yandex collector: данные Яндекс-сервисов.
- pandas: нормализация и агрегация данных.
- roi_calculator.py: ROI, cost efficiency, HITL efficiency.

CSV EXPORT

- raw_data — сырые данные по секциям.
- aggregates — агрегированные показатели.
- changelog — журнал изменений.
- report_snapshot — снимок отчёта.
- Колонки формируются динамически из ключей данных.



Диаграмма потоков данных. Аудит, семантика и отчётность.



[ВСТАВИТЬ КОД] events/publisher.py, events/consumer.py, celery_app.py



Схема базы данных — аналитический контур (PostgreSQL)

CRAWL_RESULTS

PK audit_id UUID
FK project_id
url TEXT
status_code INTEGER
title / description / h1
raw_data JSONB

PUBLIC_AUDIT_RESULTS

PK id UUID
url TEXT
audit_data JSONB
created_at TIMESTAMPTZ
expires_at TIMESTAMPTZ

SEMANTIC_RESULTS

PK id UUID
FK audit_id
ff_score FLOAT
freshness / familiarity / quality
eeat_score FLOAT

LLM_DRAFTS

PK id UUID
project_id UUID
url TEXT
draft_type TEXT
draft_content TEXT
prompt_hash TEXT

REPORTS

PK id UUID
project_id UUID
report_type TEXT
generated_at TIMESTAMPTZ
csv_path TEXT
snapshot_data JSONB

ИНДЕКСЫ / ХРАНЕНИЕ

idx_crawl_project_date
idx_url_hash
idx_semantic_project
JSONB для сырых SEO-данных
Retention public audit: 7 дней



Технологический стек

BACKEND

Python 3.11

FastAPI

Uvicorn

SQLAlchemy

Pydantic

DATA / AI

OpenAI API

Google Gemini API

NLTK / spaCy

pandas

ТЕСТИРОВАНИЕ

pytest

coverage.py

Docker-based integration tests

CRAWLING / SEO

Scrapy

Playwright

PageSpeed Insights API

Schema.org JSON-LD

ИНФРАСТРУКТУРА

PostgreSQL 14

JSONB

RabbitMQ

Redis

Celery

Docker Compose



Демонстрация приложения



Тестирование и результаты

PYTEST — АНАЛИТИЧЕСКОЕ ЯДРО

Модуль	Покрытие
ff_score_calculator.py	100%
eeat_analyzer.py	98%
roi_calculator.py	91%
metrics/calculator.py	100%
csv_exporter.py	43%

Итог: 16 unit-тестов ядра проходят локально, покрытие по тестируемым модулям — около 74%.

ИНТЕГРАЦИОННЫЕ ПРОВЕРКИ

- ✓ Audit Service health
- ✓ Semantic Service health
- ✓ Reporting Service health
- ✓ Проверка FF-Score на контрольном наборе
- ✓ Проверка Redis-кэша LLM
- ✓ Проверка CSV export
- ✓ 18 тестов требуют Docker-стек: PostgreSQL, Redis, RabbitMQ

P95 audit start
0.323s

Unit passed
16

Total tests
34



Основные результаты и выводы

ЧТО РАЗРАБОТАНО

- Audit Service — публичный и полный аудит, Scrapy + Playwright, CWV, Schema.org.
- Semantic Service — FF-Score, E-E-A-T, Content Gap, LLM drafts.
- Reporting Service — GSC/GA4/Yandex collectors, pandas aggregation, CSV export.
- Асинхронная модель — CrawlCompleted и FFScoreRecalculated через RabbitMQ/Celery.
- Data/AI-инфраструктура — PostgreSQL JSONB, Redis LLM cache, OpenAI/Gemini.

НОВИЗНА И ПРАКТИЧЕСКАЯ ЗНАЧИМОСТЬ

- Снижение зависимости от дорогих SEO-инструментов.
- Прозрачная формула FF-Score и воспроизводимый контрольный расчёт.
- Поддержка JS/SPA-аудита через Playwright.
- LLM-рекомендации без автопубликации, совместимые с HITL-контролем.
- Возможна регистрация как РИД: программа для ЭВМ с оригинальной микросервисной реализацией аналитического контура.

ТЕСТИРОВАНИЕ

Модуль	Тестов	Итог
Backend (pytest)	37	✓ coverage 42%
Frontend (Vitest)	8	✓ PASSED

ВОЗМОЖНОСТЬ ПУБЛИКАЦИИ

- Архитектурное решение HITL + HMAC-деплой может быть описано в конференционной статье
- Открытый исходный код: github.com/kAmAZ07/SEO_Master



Возможные направления дальнейшей разработки

МАСШТАБИРОВАНИЕ

- Kubernetes для воркеров при > 100K страниц
- Browserless.io кластер для Playwright
- HA: ≥ 2 реплики каждого сервиса

Инфраструктура

НОВЫЕ АДАПТЕРЫ

- Bitrix24 CMS адаптер
- Shopify адаптер
- Custom REST-адаптер
- Реальный staging-тест WP/Tilda

Интеграции

КАЧЕСТВО КОДА

- Pydantic V2 migration (устранение 69 warnings)
- Lifespan вместо on_event (deprecated)
- Frontend coverage: цель $\geq 30\%$
- Нагрузочный тест: расширить до 50+ VU на staging

Технический долг

ПРОДУКТОВОЕ РАЗВИТИЕ

- A/B-валидация FF-Score (6–12 мес.)
- Мобильное приложение для HITL
- Публичный SaaS-режим

Продукт



Список использованных источников

- Документация Ahrefs API [Электронный ресурс]. – URL: <https://ahrefs.com/api> (дата обращения: 21.05.2026).
- Brown C., Hasan A. Search Engine Optimization: Theory and Practice [Text] / C. Brown, A. Hasan. – Sebastopol : O'Reilly Media, 2022. – ISBN 978-1-492-09623-5.
- Celery Documentation [Электронный ресурс]. – URL: <https://docs.celeryq.dev> (дата обращения: 21.05.2026).
- FastAPI Documentation [Электронный ресурс]. – URL: <https://fastapi.tiangolo.com> (дата обращения: 21.05.2026).
- Fowler M. Patterns of Enterprise Application Architecture [Text] / M. Fowler. – Boston : Addison-Wesley, 2002. – ISBN 0-321-12742-0.
- OWASP. HMAC-based Authentication Guidelines [Электронный ресурс] // OWASP Foundation. – URL: https://owasp.org/www-project-heat-sheets/cheatsheets/HMAC_Based_Authentication_Cheat_Sheet.html (дата обращения: 21.05.2026). Примечание: ссылка уточнена до конкретной страницы, так как общая owasp.org слишком обширна.
- React Documentation [Электронный ресурс]. – URL: <https://react.dev> (дата обращения: 21.05.2026).
- Richardson L. RESTful Web APIs [Text] / L. Richardson, S. Ruby. – Sebastopol : O'Reilly Media, 2013. – ISBN 978-1-449-35806-8.
- Similarweb. SEO Tools Market Overview, 2024 [Электронный ресурс]. – URL: <https://www.similarweb.com> (дата обращения: 21.05.2026).
- Simon Brown. The C4 model for visualising software architecture [Электронный ресурс] / S. Brown. – 2018. – URL: <https://c4model.com> (дата обращения: 21.05.2026).
- Tilda Publishing API Reference [Электронный ресурс]. – URL: <https://help.tilda.ws/api> (дата обращения: 21.05.2026).
- WordPress REST API Documentation [Электронный ресурс] // WordPress Developer Resources. – URL: <https://developer.wordpress.org/rest-api> (дата обращения: 21.05.2026).

