Техническое задание (черновик)

Название проекта: Сервис сбора и суммаризации научных статей с публикацией в Telegram
Версия: 0.1 (черновик для согласования)
Дата: 03.11.2025
Часовой пояс админки: Europe/Amsterdam; **публикации:** Europe/Moscow (MSK)

---

1. Цели и ценность

Создать сервис, который: 1) Обходит (кроулит) заданные источники научных публикаций.
2) Забирает полные тексты/метаданные статей, сохраняет их в базе.
3) Строит краткое, понятное резюме (plain language summary) без жаргона.
4) Публикует итоговое summary (с заголовком, тегами, ссылкой на оригинал) в Telegram-канал.

Ключевые эффекты: экономия времени читателей; повышение охвата публикаций; единая база знаний с быстрым поиском.

---

2. Глоссарий и определения

- **Источник:** сайт/лента (arXiv, PubMed, журналы, блоги институтов и т.п.).
- **Запись/Статья:** единица контента (paper, preprint, editorial, dataset note и т.п.).
- **Суммаризация:** автоматическое сжатие содержания с фокусом на понятность неф специалисту.
- **Pipelines:** независимые этапы обработки: crawl → fetch → parse → extract → summarize → publish.
- **Дубль:** одна и та же статья, найденная в нескольких источниках (по DOI/URL/заголовку).

---

3. Стейкхолдеры

- Владелец продукта.
- Редакция/модераторы (опционально).
- Администраторы/DevOps.
- Читатели Telegram-канала.
- Автор(ы) статей (косвенно).

---

4. Область охвата (Scope)

Входит

- Подключение n источников.
- Планировщик обхода.
- Загрузка/парсинг статей (HTML/PDF/TeX, по возможности).
- Очистка, извлечение текста/метаданных, дедупликация.

- Автосуммаризация с настройками длины/тона.
- Публикация в Telegram, логирование и мониторинг.
- Веб-админка/панель оператора (MVP: простая).

Не входит на первых этапах (может войти позже)

- Платёжные функции.
- Обход платных стен (paywall) и скрейпинг приватного контента.
- Полноценная многоязычная локализация админки.
- Веб-портал для конечных читателей (кроме Telegram).

---

5. Пользовательские сценарии (User Stories)

- **US-1:** Как редактор, хочу добавить новый источник (RSS/сайт/arXiv категорию), чтобы сервис сам стал его мониторить.
- **US-2:** Как система, хочу по расписанию обходить источники, чтобы своевременно находить новые статьи.
- **US-3:** Как система, хочу скачивать и парсить статью (метаданные + полный текст), чтобы сохранять её в базе.
- **US-4:** Как читатель, хочу читать краткое и понятное summary, чтобы быстро понять суть работы.
- **US-5:** Как редактор, хочу согласовывать/править summary перед публикацией (опционально), чтобы контролировать качество.
- **US-6:** Как система, хочу публиковать summary в Telegram-канал с ссылкой и форматированием, чтобы обеспечить охват.
- **US-7:** Как администратор, хочу мониторить состояние pipeline и ошибки, чтобы быстро реагировать.

---

6. Функциональные требования

6.1 Управление источниками

- **FR-1:** Реестр источников с типом: RSS/Atom, sitemap, статические страницы, API (arXiv, Crossref, PubMed), кастомные парсеры.
- **FR-2:** Поля источника: `id, name, type, url(s), parser_type, rate_limit, enabled, priority, default_tags`.
- **FR-3:** Стратегии обнаружения: инкрементальный (по дате), по списку, по ключевым словам/категориям.
- **FR-4:** Настройка расписаний (cron-подобно) и ограничений на одновременные задания; **частота обхода — не чаще 1 раза в сутки на источник**.

6.1.1 Стартовые источники (из ответа 1)

Приоритет A — Пир-ревью «первым делом» (высокий приоритет публикации): - Nature (nature.com), включ. Nature Methods, Nature Biotechnology, Nature Communications; отслеживать также рубрику *News & Views* к соответствующим статьям. - Science / AAAS (science.org), раздел *First Release*. - Cell Press (cell.com) — прежде всего *Cell* и *Molecular Cell*. - PNAS (pnas.org) — *Latest / Early*

Edition. - eLife (elifesciences.org) — непрерывная публикация, строгая редакция. - Nucleic Acids Research / OUP (academic.oup.com/nar) — OA, ежегодные выпуски *Database/Servers*.

Приоритет B — Валидация и эталоны (для проверки фактов/обогащения, не первичный источник для постов): - CASP / спецвыпуски в *PROTEINS* (predictioncenter.org) — независимая оценка методов. - wwPDB / RCSB PDB (rcsb.org) — эталонные депозиты структур. - UniProt / EMBL-EBI (uniprot.org) — проверенные аннотации и последовательности. - NobelPrize.org (nobelprize.org) — первичный источник по Нобелевским премиям.

Приоритет C — Авторитетные обзоры к свежим статьям (для контекста): - Nature — *News & Views*. - Science — *Perspectives*.

Приоритет D — Препринты (раннее обнаружение; публиковать с явной пометкой #preprint и дисклеймером): - bioRxiv (biorxiv.org) — биология. - arXiv (arxiv.org) — ИИ/математика/CS и др.

Приоритет E — Оповещения/эмбарго-ленты (источник сигналов; публиковать только после официального релиза/снятия эмбарго): - Nature Briefing и тематические подборки. - PNAS Press / Embargoed media. - EurekAlert! (eurekalert.org).

Методы подключения (по источнику): RSS/Atom, официальные API (если есть), либо HTML-парсинг списков/ленты; частоты и окна обхода уточняются в Вопросе #2.

6.2 Обход и загрузка

6.2.1 Частоты обхода (утверждено)

- **Глобальное правило:** каждый источник обходится **не чаще 1 раза в сутки**.
- **Окно запуска обхода:** ночное окно по MSK (например, 03:00–05:00), настраиваемое.
- **$1-** Повторные попытки (утверждено): **если суточный обход источника завершился сбоем, выполнить** один **ретрай через** 5–10 часов** (случайный сдвиг). После этого — ждать следующего дня.
- **FR-5:** Планировщик заданий (distributed), очереди для кроулеров/фетчеров.
- **FR-6:** Соблюдение `robots.txt`, `crawl-delay`, заголовок `User-Agent`.
- **FR-7:** Rate limiting по домену/источнику; backoff при ошибках/429.
- **FR-8:** Загрузка HTML/PDF/JSON; опционально – TeX (arXiv) и ZIP с исходниками.
- **FR-9:** Встроенная обработка редиректов, HTTPS, временных ошибок, повторные попытки (idempotency).

6.3 Парсинг и извлечение

- **FR-10:** Извлечение метаданных: заголовок, авторы, аффилиации, аннотация, DOI/ArXiv ID/PMID, журнал/конференция, дата, лицензия, ключевые слова, ссылки, язык.
- **FR-11:** Извлечение основного текста (из HTML/PDF) с очисткой от навигации/рефов (по возможности).
- **FR-12:** OCR для сканов (PDF) при необходимости. **MVP: выключено по умолчанию; включается флагом** `enable_ocr`**.**
- **FR-13:** Определение языка и трансформация к целевому (напр., RU/EN) при публикации.

- **FR-14:** Дедупликация по DOI/заголовку+авторам+фингерпринт текста (shingling/SimHash/MinHash).
- **FR-15:** Нормализация ссылок на оригинал (если DOI – строить резолверную ссылку; по возможности искать открытый full-text).

6.4 Суммаризация

- **FR-16 (Длина — утверждено):** содержательная часть поста (summary) — до **1000 символов**.
- **FR-16a:** лимит **не включает** метаданные (ссылка на оригинал, авторы, DOI/журнал, теги) — они выводятся вне текста summary.
- **FR-17:** Тон: «простыми словами», избегая жаргона; сохранять смысл и ограничения исследования. $1- **FR-18a (Низкая уверенность/ошибки):** даже при низкой уверенности, частичном парсинге или ошибках извлечения система **всегда формирует черновик** для модерации.
- **FR-18b (Маркировка):** такие черновики помечаются тегом `needs_review` и подсвечиваются в UI.
- **FR-19:** Контекстные детали: что сделали, на чём проверили (n/датасет), что получили, ограничения/предостережения. $1- **FR-20a (Отображение):** Теги используются для внутренней классификации и поиска; **не выводятся** в постах.
- **FR-21:** Дисклеймеры: «preprint / без peer-review» при необходимости; отметка #preprint.
- **FR-21a (Контекст):** (Контекст): *при наличии соответствующих обзоров (Nature News & Views*, *Science Perspectives*) извлекать 1–2 ключевые мысли для контекстуализации summary (с явной ссылкой/атрибуцией в админке; в пост — опционально).
- **FR-21b (Препринты):** если источник — bioRxiv/arXiv, добавлять дисклеймер «preprint, без peer-review», пометку #preprint и (при появлении версии в журнале) обновлять «каноническую» ссылку.
- **FR-21c (Проверка фактов для белковых результатов):** по возможности соотносить упомянутые структуры/базы с RCSB PDB, UniProt; указывать идентификаторы (PDB ID, UniProt ID) и несоответствия маркировать флагом «низкая уверенность».

6.5 Публикация в Telegram

- **FR-22:** Интеграция с Telegram Bot API: отправка в заданный канал.
- **FR-23 (Формат — утверждено):** один пост = одна статья; дайджесты отключены. Шаблон: заголовок, summary (до 1000 символов), метаданные (вне лимита): ссылка на оригинал (обязательно), DOI (если есть), журнал/год **только при наличии DOI**, авторы (до **3**; при >3 — *et al.*). Эмодзи и хэштеги **не использовать**.
- **FR-23a (Модерация — утверждено):** все публикации создаются как **черновики**. Публикация происходит **только после явного утверждения** владельцем. Если черновик **не утверждён** — в этот день **ничего не публикуется**.
- **FR-23b (Доставка черновиков):** черновики направляются владельцу **в личные сообщения Telegram** (бот ↔ `owner_telegram_user_id`) и доступны в **UI проекта** (страница «Черновики»), с кнопками: *Утвердить*, *Отказать*, *Редактировать*.
- **FR-23c (Правило 11:00):** черновики, **утверждённые до 11:00 MSK**, попадают в выпуск этого дня (отправляются в 11:00). Черновики, утверждённые **после 11:00**, переносятся на **11:00 следующего дня** (если правило окна отбора не запрещает; см. FR-25).
- **FR-24 (График — утверждено):** публикация **ежедневно в 11:00 по Europe/Moscow (MSK)**. Если за прошедшие сутки нет новых **и/или утверждённых** статей — **ничего не публикуем**.
- **FR-25 (Окно отбора — утверждено):** в публикацию попадают статьи, обнаруженные **в календарные сутки, предшествующие дате постинга по MSK** (с 00:00:00 до 23:59:59

предыдущего дня). Альтернативный режим «последние 24 часа до 11:00» — опциональная настройка.
- **FR-26:** Idempotency: защита от дублей при повторах и сбоях; повторные отправки с тем же `idempotency key` не создают новые посты.
- **FR-27 (Анти-флуд):** при множестве статей отправка последовательной пачкой, начиная в 11:00, с настраиваемым интервалом (2–10 сек) между сообщениями.
- **FR-29 (Multi-channel/EN):** конфигурация каналов с привязкой к языку (RU/EN), `chat_id` и таймзоне; **EN-канал предусмотрен, но по умолчанию выключен**. Для EN допускается отдельный график (по умолчанию не задан).

6.6 Админ-панель / API оператора

- **FR-27:** CRUD источников, просмотр логов/очередей, ручной перезапуск задач.
- **FR-28:** Просмотр контента статьи, сгенерированного summary, статуса публикации.
- **FR-29:** Ручное редактирование summary, модерация и утверждение.
- **FR-30:** Поиск по базе: по DOI, авторам, заголовку, тегам, дате, источнику.
- **FR-31 (Модерация — утверждено):** раздел «Черновики» с фильтрами (источник, дата обнаружения, журнал, теги, #preprint), действиями массового выбора (утвердить/отклонить), журналом правок и кнопкой «Отправить предпросмотр в Telegram DM».
- **FR-32 (Настройки владельца):** хранение `owner_telegram_user_id`, включение/выключение DM-доставки черновиков, выбор Markdown/HTML, настройка интервалов анти-флуда.

6.7 Логи, аудит, мониторинг

- **FR-31:** Подробные структурированные логи; трассировка задач (trace id).
- **FR-32:** Метрики: время кроулинга, доля успешно распознанных PDF, средняя длина summary, время публикации, ошибки по типам.
- **FR-33:** Алерты на SLO-превышения/ошибки: **встроенные уведомления в UI** (панель событий/уведомлений). Внешние каналы (Slack/Email/Telegram) **в MVP не используются**.

---

7. Нефункциональные требования

- **NFR-1 Производительность:** обработка $\geq$ X новых статей/час при Y источниках; суммаризация $\leq$ Z сек/статья (параметры уточнить).
- **NFR-2 Масштабируемость:** горизонтальное масштабирование воркеров; очереди задач.
- **NFR-3 Надёжность:** автоматические повторы с backoff; гарантированная поставка в Telegram при сетевых сбоях.
- **NFR-4 Доступность:** целевая SLO доступности 99.5% для публикаций.
- **NFR-5 Безопасность:** хранение секретов (бот-токен, API-ключи) в vault/переменных окружения; HTTPS; ограничение прав сервис-аккаунтов; **админка за Nginx Basic Auth (логин/пароль)** для пет-проекта.
- **NFR-6 Соответствие:** соблюдение ToS источников, robots.txt, лицензий (CC-BY и т.п.), GDPR (персональные данные в авторах/аффилиациях).
- **NFR-7 Обслуживаемость:** инфраструктура как код; развёртывание через CI/CD; blue/green или canary для обновлений.
- **NFR-8 Наблюдаемость:** сбор метрик/трейсов/логов; дашборды.
- **NFR-9 Стоимость и квоты (утверждено):** минимальные затраты; **локальная LLM по умолчанию**; внешние API/LLM — опционально и по умолчанию **выключены**; разрешены только бесплатные/безключевые интеграции; троттлинг/квотирование вызовов.

- **NFR-10 Доступ к источникам: без платных подписок**; обход **paywall** не допускается; используем только открытые метаданные/аннотации/полные тексты в рамках лицензий.

---

8. Архитектура (предложение)

Компоненты: 1) **Source Registry** (реестр источников).
2) **Crawler** (обнаружение новых записей).
3) **Fetcher** (загрузка HTML/PDF/API).
4) **Parser/Extractor** (очистка, метаданные, текст).
5) **Deduplicator** (DOI + текстовый фингерпринт).
6) **Summarizer Service** (локальная open-source LLM по умолчанию; внешние LLM/API — опциональные и **выключены**; политика безопасности)..
7) **Publisher** (Telegram, форматирование, отложенный постинг).
8) **Admin UI + Operator API**.
9) **Storage/Infra (утверждено):** PostgreSQL (метаданные), файловое хранилище (NFS/POSIX) для PDF/артефактов (**без S3**), кеш — Redis, брокер задач — RabbitMQ. 10) **Observability:** графики и дашборды в Grafana; сбор метрик — по согласованию (по умолчанию Prometheus-совместимые экспортёры), логи — централизованно.

Примечания: - Idempotency keys на уровне задач; детерминированные имена файлов в файловом хранилище. - Ограничители по доменам; конфигурируемые парсеры на основе шаблонов/селекторов. - Валидация лицензий: если лицензия запрещает redistrib/quote, публикуем только summary + ссылку. - **Стек (утверждено):** Node.js/**TypeScript**; PostgreSQL; RabbitMQ; Redis; файловое хранилище (не S3).

---

9. Модель данных (черновик)

- **Article**: `id, external_id (DOI/arxiv/pmid), title, abstract, authors[], affiliations[], journal, year, volume, issue, pages, url, fulltext_url, license, language, created_at, updated_at`.
- **ArticleText**: `article_id, raw_text, cleaned_text, sections(json), ocr_used(bool)`.
- **Summary**: `id, article_id, summary_text, length_type, tone, confidence_score, tags[], warnings[], created_at, created_by(model/version)`.
- **Source**: `id, name, type, base_url, parser_config(json), schedule, rate_limit, enabled, priority, default_tags`.
- **Publication**: `id, summary_id, channel_id, status(enum: draft/approved/scheduled/sent/failed/cancelled), message_id, dm_preview_message_id, approved_by, approved_at, published_at, retry_count`.
- **Logs/Metrics**: события pipeline, ошибки, длительности.

$1 - **Научные и справочные API (для обогащения/проверки):** arXiv API; (при наличии) bioRxiv/OAI; Crossref (метаданные/DOI); RCSB PDB REST; UniProt REST; (при необходимости) Europe PMC / Entrez. - **Медийные ленты:** Nature Briefing, PNAS Press, EurekAlert! (используются как сигнальные

источники; соблюдение эмбарго). - **Подписки на журналы:** не используются; **paywall-контент не скрейпим**; работаем только с открытыми источниками/метаданными в рамках лицензий.

---

11. Правила форматирования поста (пример шаблона)

11.1 Одна статья (по умолчанию)

```
<Заголовок>

<Summary — до 1000 символов, 3–5 предложений простым языком>

Оригинал: <url>
Авторы (до 3): <Фамилия И., Фамилия И., Фамилия И., et al.>
DOI/Журнал/Год: <заполнять только если есть DOI; иначе строку опустить>
```

Примечание: метаданные **не входят** в лимит 1000 символов. Эмодзи и хэштеги не используются.

---

12. Качество summary и защита от ошибок

- Правдоподобность: только факты из текста; флаг низкой уверенности.
- Тесты на «галлюцинации»: сравнение ключевых фактов с извлечёнными разделами.
- Длина и читаемость: целевой уровень B1–B2; избегать узкого жаргона; гайд по стилю.
- Авто-детект «опасных доменов» (медицина/здоровье/био): добавлять дисклеймер «не медсовет».
- Проверка на ретракции/вывод «предварительный/preprint».

---

13. Безопасность и соответствие

- Секреты: хранить в менеджере секретов; ротация.
- Доступы по ролям (RBAC) в админке.
- Лицензии и ToS источников: не публиковать полные тексты; только summary + ссылка; уважать robots.txt.
- GDPR: хранить только публичные авторские данные; обработка запросов на удаление по требованию.

---

14. Развёртывание и эксплуатация

- Контейнеризация (Docker); оркестрация — **Docker Compose** (**без Kubernetes**).
- Окружения: dev/stage/prod; миграции БД.
- CI/CD: авто-тесты, линтеры, деплой с откатом.
- Мониторинг: Grafana (дашборды); сбор метрик — Prometheus-совместимые экспортёры; алерты в Telegram/Slack админов.
- Резервные копии БД и файлового хранилища: **ежедневно**, retention **7/30** (7 последних дневных и 30 месячных снапшотов); периодические тесты восстановления.

15. Тестирование и приёмка

- **Юнит-тесты:** парсеры, нормализация, дедупликация.
- **Интеграционные тесты:** связки crawler→fetcher→parser→summarizer→publisher.
- **E2E:** добавление источника → автопубликация в тестовый канал.
- **Нагрузочные:** X статей/час, устойчивость к пикам.
- **Критерии приёмки MVP:**
- Поддержка ≥ 3 источников (например, arXiv категории, RSS крупного журнала, PubMed по ключам).
- ≥ 90% корректное извлечение метаданных (ручная выборка N=100).
- Средняя длительность pipeline ≤ 5 мин/статья (пример).
- Автопубликация в тестовый Telegram-канал с корректным шаблоном.

---

16. План релизов (дорожная карта, черновик)

- **MVP (R1):** 3 источника, базовые парсеры, одна длина summary, ручное утверждение, публикация в один канал, базовая админка, метрики/логи.
- **R2:** мульти-длины и тона, авто-теги, OCR, дедуп по нескольким сигналам, очередь отложенных публикаций, алерты.
- **R3:** расширенная модерация, multi-channel, мультиязычность, классификация тем, анти-галлюцинационные проверки.

---

19. Зафиксированные настройки (п.11–12)

19.1 Telegram / Уведомления и публикация (утверждено) - **Owner (для ЛС с черновиками):** `@amartyn0v`
При запуске требуется определить **numeric** `owner_telegram_user_id` через /start с ботом и сохранить в настройках. - **Бот для отправки (DM + канал):** `@my_tarkov_test_bot`
Требуется **BOT_TOKEN**; права на отправку в канал; режим форматирования — **Markdown** (по умолчанию). - **Канал публикации:** `chat_id` — **TBD** (нужен идентификатор канала, куда бот имеет право писать).
- **Правила публикации:** один пост = одна статья; в 11:00 MSK; только утверждённые черновики; если нет — постов нет.

19.2 MVP-объём источников (утверждено) - **Приоритет A (peer-review, “первым делом”):** Nature (включая Methods/Biotech/Communications + News & Views), Science (First Release), Cell Press (Cell, Molecular Cell), PNAS (Latest/Early Edition), eLife, Nucleic Acids Research.
- **Приоритет D (раннее обнаружение с #preprint):** bioRxiv, arXiv.
Остальные источники и ленты подключаются позже.