

ТЕХНИЧЕСКИЙ ПЛАН

Выгрузка всех документов из ЦБД Минюста Кыргызской Республики

Структура источника, рекомендуемый формат архива и план создания возобновляемого загрузчика нормативных актов, законов и других правовых документов.

Источник: cbd.minjust.gov.kg

Дата исследования: 5 августа 2026 года

1. Результат исследования

Версия приложения: cbd-downloader 0.1.0

Основной вывод: HTML-парсер не требуется. Сайт предоставляет официальный Open Data API, который надежнее и проще обхода страниц браузером.

Базовый адрес API:

`https://cbd.minjust.gov.kg/api/v1/OpenData/`

На дату исследования API сообщает о **209 685 документах** во всех разделах.

2. Разделы базы

ID	Раздел
1	Нормативные правовые акты
6	Международное право
9	Судебные акты по вопросам законодательства
14	Правовые акты Министерства юстиции

3. Структура документа

```
Документ
├── код, раздел, вид документа и статус
├── орган принятия, дата и номер
├── название на русском и кыргызском языках
├── классификаторы и связанные документы
├── редакции
│   ├── код и дата редакции
│   ├── текст RU / KY в HTML
│   └── изображения
```

Документ может иметь несколько редакций. Русский и кыргызский тексты могут присутствовать независимо. Изображения возвращаются отдельно либо в Base64.

4. Полезные методы API

Метод	Назначение
<code>CheckAvailable</code>	Проверка доступности сервиса.

Метод	Назначение
	Версия приложения: cbd-downloader 0.1.0
GetClasses	Получение разделов базы.
GetProperties	Перечень полей и типов данных.
GetDocumentListByQuery	Создание списка документов и получение первой страницы.
GetDocumentListById	Стабильная постраничная выгрузка созданного списка.
GetDocument / GetDocuments	Метаданные одного или нескольких документов.
GetEditions	Список редакций документа.
GetEditionHtml	HTML-текст конкретной редакции.
GetEditionImage	Изображения редакции.

Официальная справка: <https://cbd.minjust.gov.kg/synopsis/ru>

5. Формат локального архива

Рекомендуется скачивать все четыре раздела, все редакции, оба языка, изображения и исходные метаданные.

```
archive/
├─ manifest.sqlite
├─ documents/
│   └─ 13443/
│       ├── metadata.json
│       └─ editions/
│           └─ 10/
│               ├── ru.html
│               ├── ky.html
│               └─ images/
```

`manifest.sqlite` хранит прогресс, контрольные суммы, ошибки и даты проверки. DOCX/PDF на первом этапе создавать не нужно: конвертация увеличивает объем и может исказить исходное оформление.

6. План разработки

1. Подготовка

Создать ветку `feature/cbd-document-downloader`. Реализовать один Python-скрипт без внешних зависимостей.

2. Пробная выгрузка

Скачать около 50 документов разных типов: действующие и утратившие силу, RU/KY, с одной и несколькими редакциями, с изображениями, международные и судебные акты.

3. Получение полного списка

Версия приложения: cbd-downloader 0.1.0

```
GetDocumentListByQuery.json
?Property=Code&PageSize=1000&PageNumber=1
```

Ответ содержит идентификатор списка, общее количество и первую страницу. Остальные страницы получать через `GetDocumentListById`. Для текущего объема потребуется около 210 легких запросов.

4. Метаданные

Получать документы через `GetDocuments` пакетами по 20–50 кодов. Сохранять оригинальный JSON со статусом, видом акта, органами, датами и классификаторами.

5. Редакции и изображения

Для каждого документа получить список редакций, скачать HTML на RU и KY, затем изображения. Отсутствующий язык отмечать в БД, а не считать ошибкой.

6. Ограничение нагрузки

Параллельность	2 запроса
Скорость	1–2 запроса в секунду
Timeout	60 секунд
Повторы	5 попыток; задержки 2, 4, 8, 16, 32 секунды

Повторять запросы при HTTP 429, 500, 502, 503 и 504. Перед запуском вызывать `CheckAvailable`.

7. Возобновление загрузки

После каждого объекта фиксировать URL, HTTP-статус, размер, SHA-256, время и число попыток. Файлы с совпадающей контрольной суммой повторно не записывать.

8. Проверка полноты

- Число уникальных кодов равно `TotalCount`.
- Нет дубликатов.
- У каждого документа сохранены метаданные.
- Все заявленные редакции и изображения обработаны.
- SHA-256 сохраненных файлов совпадает.
- Редакции и языки сверены с sitemap.

9. Инкрементальные обновления

Sitemap содержит URL редакции, язык и `lastmod`. При повторном запуске достаточно скачать новые или изменившиеся записи, периодически выполняя полную сверку кодов.

7. Минимальная реализация

Достаточно стандартной библиотеки Python:

```
urllib.request · json · sqlite3 · pathlib · hashlib · gzip  
xml.etree.ElementTree · concurrent.futures
```

Версия приложения: cbd-downloader 0.1.0

Предлагаемый интерфейс:

```
python download_cbd.py --output archive  
python download_cbd.py --resume  
python download_cbd.py --latest-only  
python download_cbd.py --classes 1,6,9,14  
python download_cbd.py --verify
```

Первая версия — **0.1.0**. Итоговая строка каждого запуска показывает версию, число скачанных документов, ошибок и пропусков.

8. Источники

- [ЦБД Минюста Кыргызской Республики](#)
- [Официальная справка и описание Open Data API](#)
- [Основной sitemap документов](#)
- [robots.txt](#)