Audion Office OCR AI

English · Руководство · История

Содержание

Распознавание сканов и офисная конверсия: из бумаги в редактируемый документ, таблицу, проверяемый PDF и всё остальное — за одно распознавание.

Зачем это сделано

Распознавание стоит денег и времени. Обычный порядок работы этого не учитывает: прогнали скан, получили Word — а через неделю понадобилась таблица. Прогнали заново. Понадобился PDF с текстовым слоем — ещё раз.

Каждый прогон — новые деньги, новая очередь и новый результат, который может отличаться от прежнего.

Здесь порядок другой:

```
источники → распознавание → пакет документа → проверка → любые форматы
```

Платите за распознавание один раз. Результат складывается в независимый пакет `<имя>.document`, и дальше все форматы собираются из него локально — без обращения к платной службе.

Что в пакете

Не текст, а всё, из чего он получился:

- точная копия исходника;
- изображения всех страниц до и после подготовки;
- хэши;
- текст и координаты **всех** кандидатов распознавания, а не только выбранных;
- порядок чтения, таблицы, объединения ячеек;
- уверенность по каждому фрагменту;
- сверка между разными движками;
- сырой ответ поставщика;
- место для ручных исправлений.

Полнота здесь означает отсутствие потерь исходных данных, а не обещание безошибочного распознавания. Ошибка распознавания видна и исправима, потому что рядом лежит всё, на чём она основана.

Что собирается из пакета

Восемь результатов, независимо и одновременно: DOCX, XLSX, PDF с текстовым слоем, ODT, Markdown, JSON распознавания, HTML и лист проверки.

По умолчанию собирается только DOCX — форматных наборов нет, лишнее не производится. Пересборка идёт локально, сторонний офисный пакет для неё не нужен. Предельный размер страницы — A3.

Принципы

Второй проход другим движком — независимый. Он получает тот же очищенный растр, но не видит текста первого. Так расхождение между ними означает реальную неоднозначность, а не подсказку, повторённую следом.

Таблицы разбираются по физической сетке. Для линованных таблиц строится сетка по длинным линиям растра, а число столбцов проверяется по строке нумерации `1…N` из координат распознанных слов. И только потом движки сравниваются внутри точных ячеек.

Объединения восстанавливаются **только в шапке** — там они предсказуемы. В теле таблицы объединённая ячейка чаще означает ошибку разбора, чем замысел автора.

Расхождения цифр сохраняются. Если движки прочли число по-разному, оба варианта остаются в пакете. Число — то место, где ошибка распознавания дороже всего, и решать её должен человек.

Markdown — не центр системы. Он удобен как контрольный формат и как вход для языковых моделей, но пакет документа устроен не вокруг него.

Дальше

- Руководство — работа по шагам.
- История — что менялось.

Техническая часть

Названия в рабочей области

Строки путей называются **Источник** и **Назначение**. Кнопки те же, что во всех проектах Audion: **Источник**, **Добавить файл…**, **Назначение**, **Сбросить**, **Удалить**, **Список**.

Источник может быть отдельным файлом или папкой. Внешний путь зеркалируется во внутреннюю рабочую папку, результаты синхронизируются в выбранное назначение.

Второй проход

значение	что делает
нет	только основной движок
Tesseract	локальная сверка
Yandex	облачная сверка
Yandex и Tesseract	обе

Полный архив

Сборка всего пакета в один архив осталась возможностью движка ради совместимости, но среди основных кнопок не показывается: она нужна редко, а места занимает много.