

Технический справочник

[English](#) · [Руководство](#) · [ОКТМО](#)

Содержание

- [Разбор адреса](#)
- [Слоты](#)
- [Проходы сопоставления](#)
- [Настройки](#)
- [Источники](#)
- [Артефакты прогонов](#)
- [Оформление результата](#)
- [Устройство проекта](#)

Устройство разбора адреса, слоты, проходы сопоставления, файлы настроек и артефакты прогонов.

Разбор адреса

Сырая строка адреса раскладывается на составляющие. Порядок разбора важен: сперва отсекается хвост с номером дома, затем определяется тип и имя улицы или территории, затем населённый пункт и вышестоящая административная единица.

Что распознаётся

Типы улиц. улица, переулок, проспект, бульвар, тракт, шоссе, тупик, проезд, аллея, линия, набережная, площадь, квартал, микрорайон, километр, автодорога — со всеми обычными сокращениями (**ул.**, **пер.**, **пр-т**, **б-р**, **наб.**, **мкр.**).

Типы территорий. тер., территория, СНТ, ДНТ, ТСН.

Типы населённых пунктов. город, посёлок, деревня, село, слобода, станица, станция, хутор, улус, местечко, кишлак, аул, аал, арбан, починок, выселок, заимка, кордон, маяк, погост, слободка, усадьба, лесоучасток, метеостанция, разъезд. Отдельно разбираются железнодорожные формы: `п.`, `ж/д ст.`, `ж/д остановочный пункт`, `ж/д блокпост`, `ж/д будка`, `ж/д ветка`, `ж/д казарма`, `ж/д платформа`, `ж/д площадка`, `ж/д путевой пост`, `рзд.`. Компактные формы `гп`, `рп`, `кп`, `дп`, `пгт`, `нп` понимаются с точками и без.

Номер дома. Хвост строки разбирается на базу и модификаторы: `10`, `10а`, `10/2`, `10к2`, `10` `корп. 2`, `10 стр. 1`, `10 литера А`. База — число, модификаторы — всё остальное, приведённое к одному виду.

Ключевое правило номера дома

Модификаторы дома нельзя схлопывать с чистым номером. `10к2` не забирает `10`, и `10` не забирает `10к2` — ни на строгом проходе, ни на мягком. Мягкость проходов касается написания улицы, а не корпусов и литер.

Слоты

Слот — это одна именованная составляющая адреса. Из включённых слотов собирается адресная строка; в разложенном виде каждый слот становится колонкой.

Слоты адресной строки

Слот	Колонка	Короткая	Что это
<code>postal_index</code>	<code>Postal_Index</code>	<code>Index</code>	Почтовый индекс
<code>federal_district</code>	<code>Federal_District</code>	<code>Fed_District</code>	Федеральный округ
<code>parent_subject</code>	<code>Parent_Subject</code>	<code>Parent</code>	Вышестоящий субъект
<code>subject</code>	<code>Subject</code>	<code>Subject</code>	Субъект федерации
<code>autonomous_okrug</code>	<code>Autonomous_Okrug</code>	<code>AO</code>	Автономный округ
<code>municipality</code>	<code>Municipality</code>	<code>Municipality</code>	Муниципальное образование
<code>locality</code>	<code>Locality</code>	<code>Locality</code>	Населённый пункт
<code>territory</code>	<code>Territory</code>	<code>Territory</code>	Территория, СНТ и подобное

Слот	Колонка	Короткая	Что это
<code>microdistrict</code>	<code>Microdistrict</code>	<code>Mkr_Qtr</code>	Микрорайон, квартал
<code>street</code>	<code>Street</code>	<code>Street</code>	Улица
<code>house</code>	<code>House</code>	<code>House</code>	Дом целиком
<code>premise</code>	<code>Premise</code>	<code>Premise</code>	Помещение, квартира

Технические слоты

Слот	Колонка	Короткая	Что это
<code>oktmo_code</code>	<code>OKTMO_Code</code>	<code>OKTMO</code>	Код ОКТМО
<code>oktmo_name</code>	<code>OKTMO_Name</code>	<code>OKTMO_Name</code>	Наименование по ОКТМО
<code>house_base</code>	<code>House_Base</code>	<code>House_No</code>	Номер дома без модификаторов
<code>house_mods</code>	<code>House_Mods</code>	<code>House_Mods</code>	Модификаторы в приведённом виде
<code>house_modifiers</code>	<code>House_Modifiers</code>	<code>House_Parts</code>	Модификаторы списком
<code>street_numbers</code>	<code>Street_Numbers</code>	<code>Street_Nums</code>	Числа в названии улицы
<code>territory_key</code>	<code>Territory_Key</code>	<code>Territory_Key</code>	Ключ территории
<code>match_key</code>	<code>Match_Key</code>	<code>Match_Key</code>	Ключ сопоставления

Умолчания

Порядок сборки адресной строки по умолчанию — порядок слотов адресной строки из таблицы выше, сверху вниз.

Включены по умолчанию все слоты, кроме `microdistrict`: микрорайон и квартал в российских адресах чаще дублируют улицу, чем дополняют её. Колонка микрорайона при этом остаётся — из строки исключается только значение.

Заголовки слот-колонок пишутся в полном или коротком виде, это выбирается на экране обработки эталона.

Проходы сопоставления

Движок сопоставления — `Ultimate_GT_Aligner`. Для каждой адресной колонки строится пул кандидатов, и строки эталона забирают из него пары.

Найденная пара немедленно изымается из пула. Один кандидат не может быть выдан дважды, а проход, отработавший раньше, не переигрывается более поздним.

Проходы идут строго по возрастанию мягкости.

Проход	Название	Условие совпадения
0	Исходная строка	Только в режиме сбора до сопоставления: кандидат происходит из той же строки источника, и его составляющие сходятся
1	Точная строка	Совпадение приведённых текстов адреса
2	Хэш составляющих	Совпали территория, слова улицы, числа улицы, номер дома и модификаторы
2R	Мягкий хэш	То же без модификаторов в ключе — но проверка дома всё равно требует совпадения модификаторов
2C	Ключевой хэш	Только в режиме сбора: улица и дом без территории. Требуется строгая проверка дома и совместимость территориальных кодов
3	Приблизительный, внутри территории	Похожесть названия улицы внутри одного территориального кластера, порог 85
4	Приблизительный, между территориями	То же по всему пулу, порог 85, с весами
5	Мягкий	Порог 80, модификаторы дома по-прежнему защищены

Точными считаются проходы 0, 1, 2, 2C, 3 и 4. **Мягкими** — 2R и 5. Сводка в журнале разделяет их отдельными строками.

Веса на приблизительных проходах

При равной похожести названия улицы кандидат получает надбавки:

- совпал номер дома — плюс 50;
- совпала территория — плюс 30;
- совпали модификаторы дома (проход 5) — плюс 20;

- совпал территориальный контекст ОКТМО — надбавка от профиля области поиска.

Пара с несовместимыми территориальными кодами отбрасывается до подсчёта весов.

Строки без улицы

Если в эталонной строке номер дома есть, а улицы нет, приблизительное сравнение названий бессмысленно. Такие строки ищут кандидата с тем же номером дома, модификаторами и числами улицы, и выбирают лучшего по территориальным весам.

Настройки

config/project.yaml

```
address_aligner:
  city: "тюмень"           # город, вычищаемый из ключей при нормализации
  ground_truth_column: "B" # эталонная колонка; "" = автоопределение
  target_columns: "C,D"    # адресные колонки; "" = авто
  companion_mode: "auto"   # auto | none | left | right | between
  max_auto_align_columns: 5 # сколько колонок брать справа в авто-режиме

safety:
  cleanup_managed_workspace_only: true # чистить только папки проекта
  never_delete_original_inputs: true   # не удалять оригиналы исходников
```

Значения из окна имеют приоритет над файлом: пустое поле в окне означает «взять из настроек или определить автоматически», заполненное — переопределяет.

config/gui_settings.yaml

```
gui:
  language: "en"           # ru | en
  theme: "code_dark"
  emoji: false
  allow_runtime_switching: true
  advanced_open: false     # раскрывать блок «Дополнительно» сразу
  source_path: ''
  destination_path: ''
```

Прочие файлы настроек

Файл	Что в нём
<code>config/tool_manifest.yaml</code>	Описание всех команд, полей, подсказок и служебных функций
<code>config/version.json</code>	Версия и язык корневого README
<code>config/path_history.json</code>	История и закрепления путей источника и назначения
<code>config/ui_colors.yaml</code>	Цвета интерфейса
<code>config/oktmo_current_keys.txt</code>	Ключи поиска ОКТМО
<code>config/oktmo_*_pins.json</code>	Закреплённые регионы и муниципалитеты

`tool_manifest.yaml` — то, из чего строится окно. Каждая команда описана названием, описанием, набором полей и ссылкой на функцию сервисного слоя. Оба языка живут в одном файле парами `label` / `label_ru`.

Источники

Поддерживаются `.xlsx`, `.docx`, `.odt`, `.txt`, `.csv`, `.md`, `.markdown` и `.pdf` с текстовым слоем. PDF без текстового слоя не распознаётся — это не задача этой программы.

Старые `.doc` и `.xls` обрабатываются только после конвертации отдельной командой.

Временные файлы Office (`~$*`) пропускаются и удаляются при проверке исходников.

Двухуровневые шапки

Строка, которая не содержит значений-данных и стоит под горизонтально объединённым родительским заголовком, считается строкой подзаголовков. Её значения присоединяются к родительским заголовкам, а сама строка не попадает в данные.

Столбец «Проектная мощность, мест» над «по корпусам» становится «Проектная мощность, мест по корпусам» вместо безымянного `column_5`.

Артефакты прогонов

Результаты

Префикс	Что это
AddressCollection_*.xlsx	Книга собранных адресов
AddressReference_*.xlsx	Эталонный справочник

Имена содержат метку времени `ГГГГММДД_ЧЧММСС`.

Отчёты в `report/`

Файл	Прогон
address_collection_summary.json	Сбор адресов
address_collection_benchmark.json	Сверка сбора с контрольной колонкой
address_alignment_summary.json	Сопоставление адресов
reference_normalization_summary.json	Нормализация эталона
reference_processing_summary.json	Обработка эталона
table_comparison_summary.json	Сравнение двух таблиц
safe_table_join_summary.json	Перенос колонок по адресу
legacy_office_conversion_report.json	Конвертация DOC/XLS
rosstat_oktmo_update_summary.json	Обновление реестра ОКТМО

Часть операций дополнительно пишет `<имя результата>.summary.json` рядом с книгой.

Листы аудита

Перенос колонок по адресу добавляет в результат лист `AUDIT_ADDRESS_JOIN`. Сбор адресов при включённой опции пишет лист отбраковки.

Оформление результата

Книги результата проходят постобработку: ширина колонок подбирается по содержимому, длинный текст переносится, высота строк растёт, статусные колонки получают цветовую заливку.

Это не косметика: без неё каждую выгрузку приходится доводить руками, а строку с длинным адресом невозможно прочитать в узкой колонке.

Устройство проекта

Слой	Где	За что отвечает
Окно	<code>system_core/ui_nicegui/</code>	Интерфейс, рабочие папки, журнал, артефакты
Сервисы	<code>system_core/services/</code>	Связь окна с движками, отчёты, параметры
Адресный движок	<code>system_core/address_engine/</code>	Разбор, слоты, ОКТМО, справочник, сбор
Сопоставление	<code>system_core/Ultimate_GT_Aligner.py</code>	Проходы сопоставления и перенос связанных колонок
Перенос по адресу	<code>system_core/safe_table_join.py</code>	Сравнение таблиц и перенос колонок через Excel
Ядро	<code>system_core/core/</code>	Задания, пути, кодировки, манифест

Подробная карта файлов — [tools/PROJECT_FILES_GUIDE_RU.md](#).