

Алгоритмическое и программное обеспечение интеграции схожих по структуре табличных документов

Декомпозиция бизнес-процессов

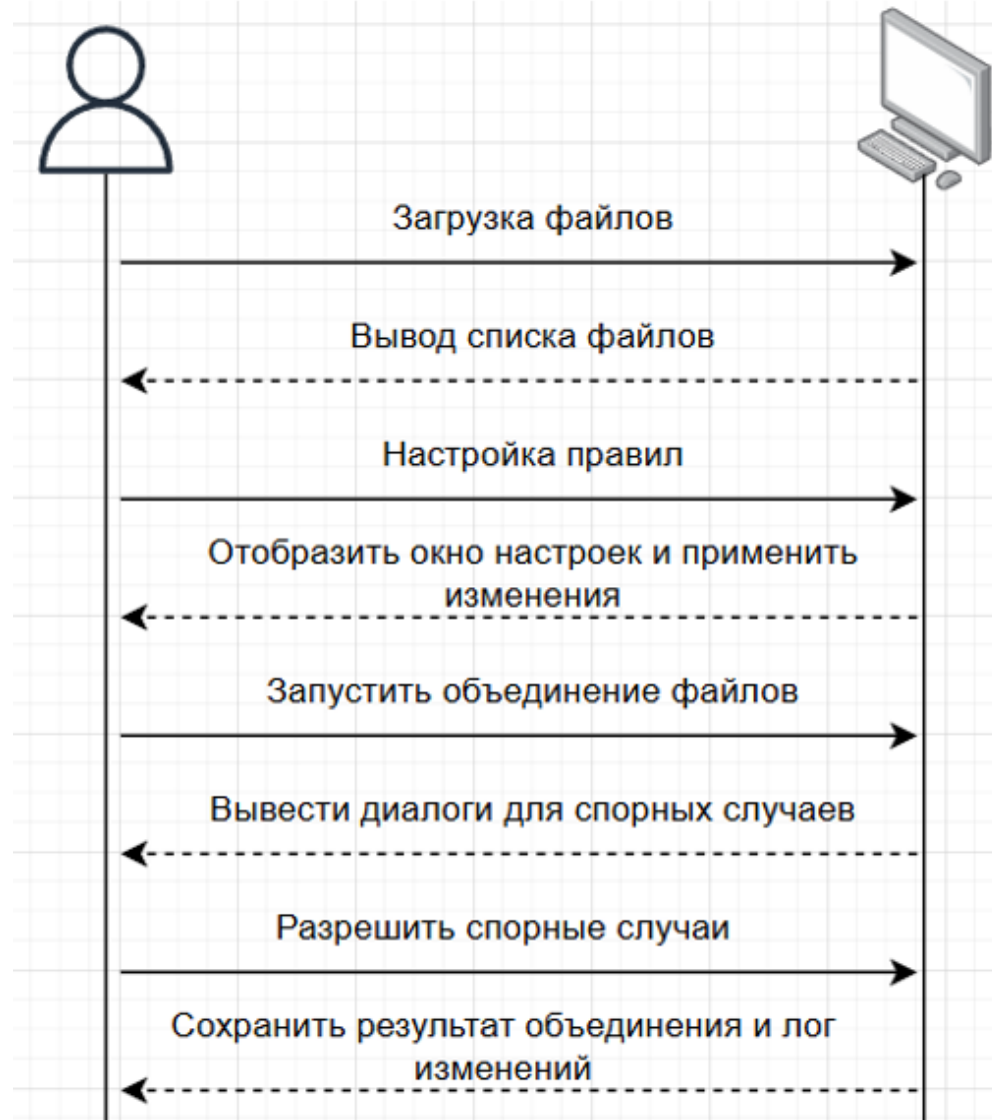
AS-IS (текущий процесс):

- Получение исходных файлов в формате Excel
- Ручная проверка заголовков и форматов дат
- Переименование столбцов в каждом файле
- Конвертация единиц вручную
- Объединение всех таблиц с помощью копирования/вставки
- Сохранение итогового файла и отправка на проверку

TO-BE (автоматизированный процесс):

- Пользователь через интерфейс загружает необходимые файлы
- Читаются файлов модулем `reader.py` и обнаружение заголовки
- На основе правил в `rules.json` и алгоритма семантического объединения `transformer.py` сопоставляются столбцы
- Автоматически приводятся даты и конвертирует единицы измерения согласно правилам в `rules.json`
- Пользователь разрешает неоднозначные случаи через диалоги
- Сохраняются файла в Excel формате и генерация лога изменений через `writer.py`

Функциональная модель взаимодействия пользователя с интерфейсом



Методы и алгоритмы решения задачи

Нечёткое сравнение «rapidfuzz»:

- поиск опечаток и синонимов в заголовках и ячейках

Семантический анализ:

- определение эмбеддингов заголовков через библиотеку «spaCy» и модель «ru_core_news_lg»
- вычитывание косинусной близости
- анализ PROPN-сущностей в содержимом
- вычитывание коэффициента Жаккара

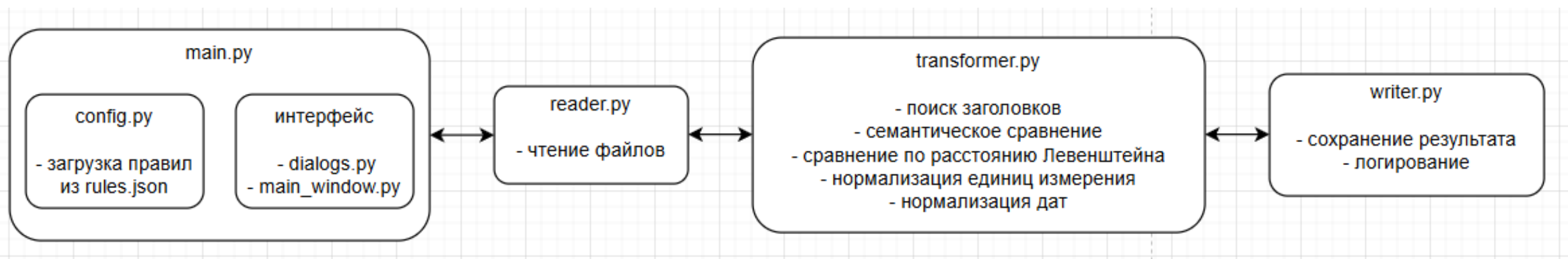
Конвертация единиц измерения:

- определение ячейки «число + единица»
- разделение и конвертация единиц измерения по словарю JSON

Настройка:

- параметры (пороги, веса, словари) настраиваются в словаре JSON без правки кода

Архитектура системы



Стек программных технологий

Необходимое программное обеспечение:

Python 3.8 – язык программирования для реализации функционала

VS Code – редактирование кода, запуск приложения на этапе разработки

Необходимые библиотеки:

PySide6 - создание интерфейса, pandas - обработка и анализ таблиц

rapidfuzz – сравнение слов с помощью расстояния Левенштейна

spacy 3.5 – семантическое сравнение строк

openpyxl и xlrd - чтение файлов в форматах xls/xlsx

re - работа с данными

datetime - приведение дат к единому формату

pandas – работа с табличными документами

logging – запись ошибок во время запуска или работы программного обеспечения

Информационная модель процесса интеграции

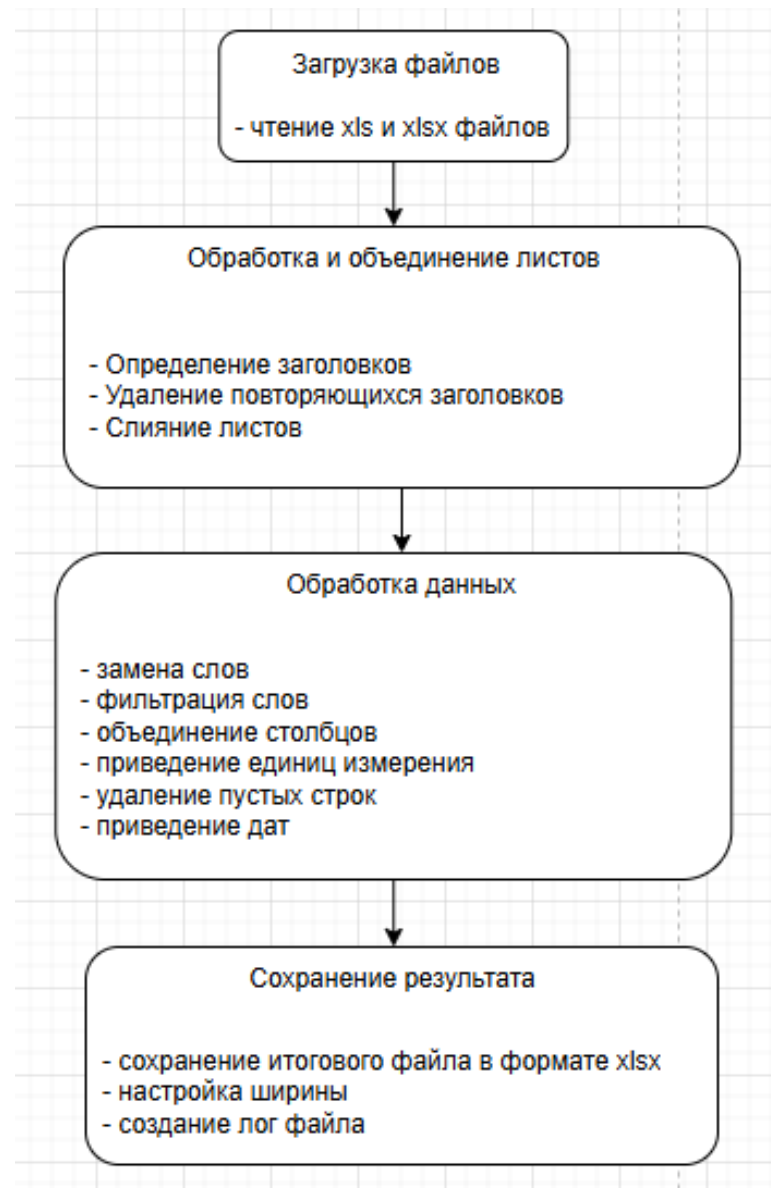
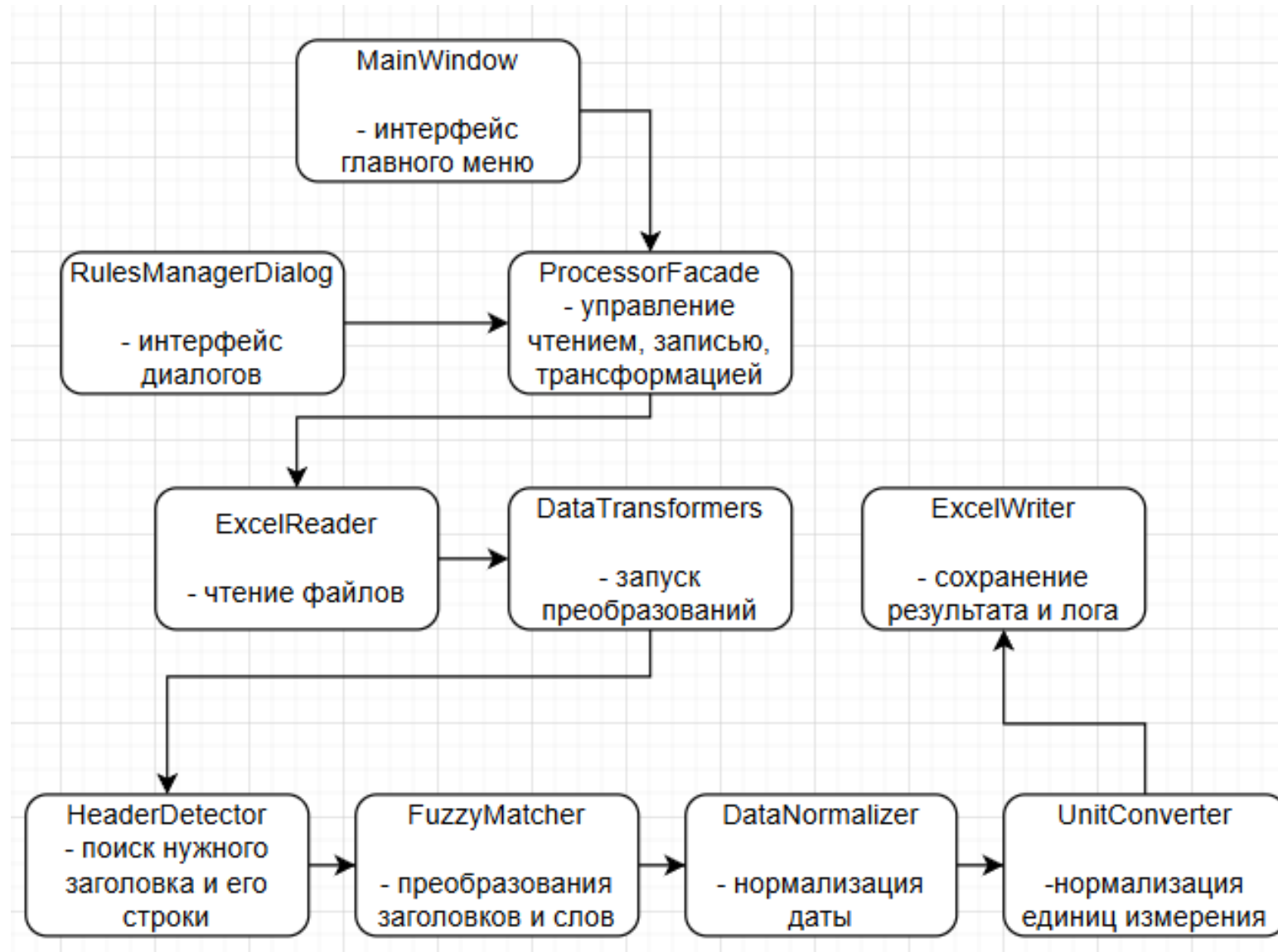


Схема взаимосвязей компонентов



Пример использования. Содержимое файлов

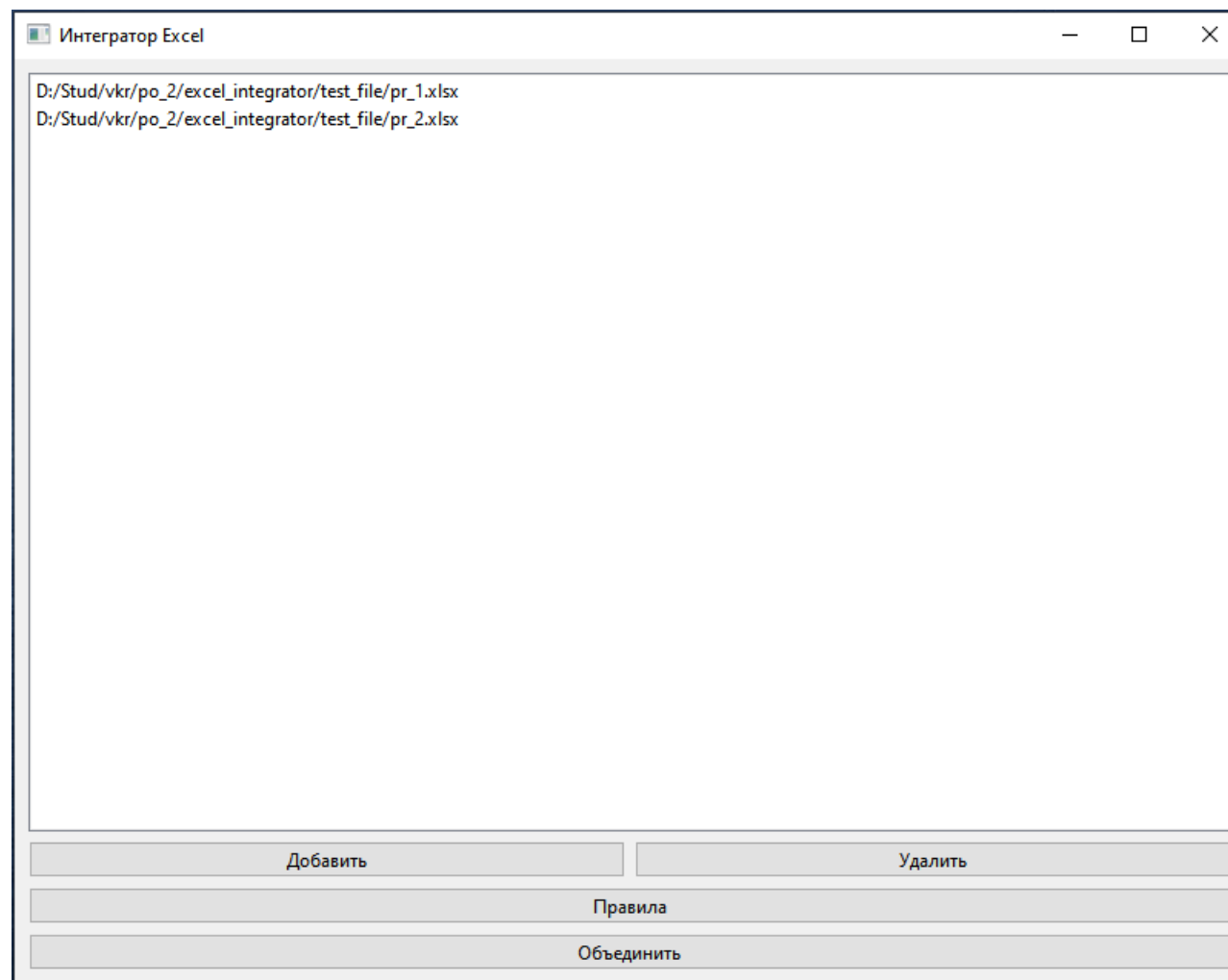
Файл 1:

	A	B	C	D	E	F
1	Локация	Дата укуса	Время укуса	Пол клеш	Свертыва	вакц-я
2	Улан-Удэ	05.08.2016	8:15:00	Самка1	20мл	да
3	Слюдянка	06.08.2016	17:50:00	Самец	20л	нет
4	Байкальск	07.08.2016	12:30:00	женский1	л20	ожидание
5						

Файл 2:

	A	B	C	D	E	
1	Место обнаружения	Свертываемость	Дата обнаружения	Пол	Вакцинация	
2	Улан-Удэ	1 л	10 августа 2016 г.	женский	есть	
3	Новосибирск	10мл	06.08.2016	мужской	неизвестно	
4						
5	Байкальск	1мл	07.08.2016	женский	есть	
6						
7	Место обнаружения	Свертываемость	Дата обнаружения	Пол	Вакцинация	
8	Слюдянка	50мл	7/3/2016	мужской	есть	
9						
10	Примечание			Калькулятор		
11	Колба	Тара 1		1		
12	Ваза	Тара 2				

Пример использования. Интерфейс главного окна



Пример использования. Интерфейс настроек и диалогов

Столбцы

	Целевое	нимы (через запя	Не объединять
1	Вакцинация	Вакц-...	<input type="checkbox"/>

Добавить правило

Удалить правило

☐ Отключить

Настройки NER

Порог схожести заголовков (0–100):

60

☐ Объединять автоматически

☒ Анализ содержимого

Число строк для анализа содержимого:

5

Вес влияния заголовков (0-1):

0,20

Сохранить

Отмена

Объединить столбцы?

☒ Объединить 'Локация' + 'Место обнаружения' (62%)

Имя результирующего столбца: Локация

OK

Cancel

Пример использования. Результат


Основной лист:

	A	B	C	D	E	F	G
1	Локация	Дата укуса	Время укуса	Пол клеща	Свертываемость	Вакцинация	
2	Улан-Удэ	05.08.2016	08:15:00	женский	20 мл	есть	
3	Слюдянка	06.08.2016	17:50:00	мужской	20000 мл	нет	
4	Байкальск	07.08.2016	12:30:00	женский	20000 мл		
5	Улан-Удэ	10.08.2016		женский	1000 мл	есть	
6	Новосибирск	06.08.2016		мужской	10 мл	неизвестно	
7	Байкальск	07.08.2016		женский	1 мл	есть	
8	Слюдянка	03.07.2016		мужской	50 мл	есть	
9							

Лист примечаний:

	A	B	C	D	E	F
1	Место обнаружения	Свертываемость	Дата обнаружения	Пол	Вакцинация	
2	Примечание			Калькулятор		
3	Колба	Тара 1		1		
4	Ваза	Тара 2				
5						
6						

Пример использования. Лог изменений

 new.log – Блокнот

Файл Правка Формат Вид Справка

Чтение файла pr_1.xlsx

Использован первый ряд как заголовок: ['Локация', 'Дата укуса', 'Время укуса', 'Пол клеща', 'Свертываемость', 'вакц-я']

Лист «2025» → «2025»

Использован первый ряд как заголовок: ['Проверка']

Лист «Sh21» → «Sh21»

Чтение файла pr_2.xlsx

Использован первый ряд как заголовок: ['Место обнаружения', 'Свертываемость', 'Дата обнаружения', 'Пол', 'Вакцинация']

Лист «2025» → «2025»

Использован первый ряд как заголовок: ['Проверка']

Лист «Sh2» → «Sh2»

Пользователь подтвердил слияние 'Sh2' → 'Sh21'

Объединение 2 частей листа «2025»

Замена слов (fuzzy): 'Самка1' → 'женский' в [0, 'Пол клеща'] (91%)

Замена слов: 'Самец' → 'мужской' в [1, 'Пол клеща']

Замена слов (fuzzy): 'женский1' → 'женский' в [2, 'Пол клеща'] (93%)

Замена слов: 'да' → 'есть' в [0, 'вакц-я']

Фильтр слов: очищена ячейка [2, "вакц-я"]

Словарно объединены ['вакц-я', 'Вакцинация'] → 'Вакцинация'

NER объединены 'Локация' + 'Место обнаружения' → 'Локация' (62%)

NER объединены 'Дата укуса' + 'Дата обнаружения' → 'Дата укуса' (66%)

NER объединены 'Пол клеща' + 'Пол' → 'Пол клеща' (72%)

Единицы в ячейках столбца 'Свертываемость': форматирование значений с единицей 'мл'

Объединение 2 частей листа «Sh21»

Перенесены строки в лист «2025_Примечание»