

Инструкция пользователя системы ОКО

Краткое руководство по загрузке, обработке, поиску и сравнению файлов утечек данных

Примечание. Инструкция подготовлена по предоставленному описанию сервиса, документации и транскрипции обучения. Скриншоты интерфейса не включены, так как в исходных материалах отсутствуют проверенные изображения интерфейса ОКО.

1. Назначение системы

Сервис ОКО предназначен для обработки и анализа файлов утечек данных. Система позволяет загружать файл утечки, привязывать его к компании, выполнять предпросмотр и валидацию данных, запускать парсинг, получать статистику, сравнивать утечки между собой и выполнять поиск строк по известным значениям.

2. Основные возможности

Возможность	Что делает пользователь
Загрузка утечки	Выбирает существующую или новую компанию, задает название утечки, указывает источник и загружает файл.
Предпросмотр данных	Проверяет первые строки файла, наличие заголовков, разделитель и предложенные системой типы столбцов.
Валидация	Смотрит, как данные приведены к единому виду, и корректирует типы столбцов при необходимости.
Парсинг и анализ	Выбирает группы данных для подсчета уникальности и сравнения строк внутри утечки и между утечками.
Поиск по утечкам	Ищет строки по значению, частичному совпадению или ключевому слову и выгружает результаты.

3. Общий порядок работы

1. Откройте раздел загрузки новой утечки.
2. Выберите компанию из списка или укажите новую компанию.
3. Укажите название утечки и источник, если он известен.
4. Загрузите файл утечки.
5. Перейдите в предпросмотр и проверьте, как система прочитала данные.
6. Выберите разделитель: запятая, табуляция, точка с запятой или свой символ.
7. Проверьте наличие заголовков. Если заголовков нет, столбцы будут обозначены номерами.
8. Проверьте автоматически определенные типы данных и при необходимости измените их вручную.
9. Посмотрите валидированный вид таблицы.
10. Настройте группы сравнения и сохраните конфигурацию.
11. Запустите парсинг.
12. Изучите статистику, результаты сравнения и при необходимости выгрузите данные.

4. Предпросмотр и валидация

В предпросмотре отображается часть файла, достаточная для первичной проверки структуры. Система предлагает типы данных для столбцов, но пользователь может изменить их вручную. По материалам обучения поддерживались следующие типы данных: электронная почта, телефон, дата рождения, паспорт, ИНН, СНИЛС, ФИО и части ФИО.

Валидированный вид показывает данные в том виде, в котором они будут храниться и использоваться при анализе. Например, значения приводятся к единому формату, а данные, не соответствующие выбранному типу столбца, не попадают в валидированный вид.

5. Группы сравнения и анализ

Группа сравнения — это набор полей, по которому система считает уникальные значения и ищет сходство между строками. Например, можно создать группу из электронной почты и даты рождения или группу из паспорта и телефона.

Важно: при строгом сравнении строка учитывается в группе только тогда, когда в ней заполнены выбранные поля этой группы. Если одно из полей отсутствует, такая строка не участвует в подсчете по данной группе.

После парсинга система показывает внутреннюю статистику по полям и группам: количество заполненных и уникальных значений. Также доступно попарное сравнение между утечками по заданным группам.

6. Поиск по утечкам

В разделе поиска пользователь может построить запрос по одному или нескольким полям. Доступен поиск по полному значению, частичному значению и ключевому слову среди данных. Результаты отображаются в виде найденных строк и могут быть выгружены в файл.

При поиске телефона система приводит номер к валидированному виду. На обучении показывался пример, когда номер, введенный через 8, распознавался как номер телефона и приводился к формату поиска.

7. Выгрузка результатов

- выборка уникальных строк утечки;
- строки, схожие с другой утечкой по выбранной группе сравнения;
- результаты поиска по значению или ключевому слову.

8. Короткий чек-лист пользователя

- Проверить, правильно ли выбраны компания, название и источник утечки.
- Проверить разделитель и наличие заголовков.
- Проверить типы столбцов до запуска парсинга.
- Открыть валидированный вид и убедиться, что данные не потерялись из-за неверного типа столбца.
- Задать группы сравнения только по тем полям, которые действительно нужны для анализа.
- После парсинга проверить заполненность и уникальность по каждому важному полю.
- При сравнении утечек убедиться, что выбраны нужные утечки для сопоставления.
- При необходимости выгрузить уникальные строки, совпадения или результаты поиска.

9. Особенности, которые нужно учитывать

- Если в одном поле файла находится сразу несколько разных значений, система может не выделить их автоматически.
- Если ФИО записано по-разному, например полностью и с инициалами, такие записи могут считаться разными.

- Частично заполненные строки сохраняются, но участвуют только в тех расчетах, для которых хватает выбранных полей.
- Перед повторным анализом убедитесь, что сохранена актуальная конфигурация типов данных и групп сравнения.