ОПИСАНИЕ ФУНКЦИОНАЛЬНЫХ ХАРАКТЕРИСТИК «ФАКТОРА»

Модули очистки и стандартизации распознают, обрабатывают и приводят к стандартному представлению адреса, телефоны, ФИО, email, даты, реквизиты компаний, документы и модели автомобилей.
Модули используют базы знаний в несколько миллионов единиц и уникальные алгоритмы для обработки данных, которые учитывают опечатки, неполноту, дублирование и разный формат данных.
Ниже для каждого модуля описаны возможности, алгоритмы, ограничения, используемые нормативные акты и справочники.

ПОЧТОВЫЕ АДРЕСА

- Разбиваем адреса на компоненты.
- Исправляем опечатки и сокращения.
- Приводим компоненты к единому формату.
- Переводим с латиницы на кириллицу.
- Восстанавливаем пропущенные компоненты: «197379 УЛИЦА ДОЛГООЗЁРНАЯ 12» «197373, г Санкт-Петербург, ул Долгоозерная, д 12»
- Заменяем старые названия на новые: «Ленинград» «Санкт-Петербург». По запросу можем оставлять устаревшие значения.
- Понимаем адрес, если регион записан в виде цифрового кода: «05 Мажалис Батырая дом 14» «Респ Дагестан, село Маджалис, ул Батырая, д 14»
- Разбираем адрес, когда тип и наименование склеились: «МОСКВА УЛКРАСНОГОМАЯКА Д, 10» «г Москва, ул Красного Маяка, д 10»
- Используем набор городов «по умолчанию», если в адресах пропущен город, но при этом известно местоположение клиентов (например, Москва и Мурманск).
- Используем ФИАС и собственные справочники: дополнения ФИАС, распространённые улицы.
- Выявляем ошибки в ФИАС. Очевидные исправляем, об остальных сообщаем в налоговую.
- В случае неоднозначности предлагаем несколько вариантов адреса.
- Почтовые индексы ФИАСа проверяем и исправляем по данным Почты России.
- Определяем:

  код КЛАДР и код ФИАС;

  guid ФИАС для всех уровней адреса;

  численность населения;

  код ИФНС для юрлиц и физлиц из ФИАС;

  признак состояния дома из ФИАС;

  район внутри города. Район внутри города также учитываем при разборе адреса, даже если его нет в ФИАС. Например, для Москвы, ни одного района которой нет в ФИАС, адрес «Москва, Хамовники р-н, ул Остоженка» найдётся и получит статус «гарантированный»

  ОКТМО и ОКАТО по адресу.

- Проставляем коды качества для распознанного адреса:

  код полноты адреса;

  статус проверки;

  признак актуальности;

  код качества сделанных изменений;

  код соответствия ФИАС;

  наполнение адреса по ФИАС.

- Записываем адреса в формате для Почты России.

- Разделяем адреса, пригодные для курьерской доставки и для почтовой рассылки. Отдельные коды полноты присваиваются адресам для доставки писем, но по которым человека не найти лично.

Ограничения

- Сопоставляем с ФИАСом до уровня улицы. Если дома нет в ФИАС, то он будет разобран, потому что ФИАС не полон по домам. Для проверки, есть ли дом в ФИАС, проставляем код соответствия ФИАС.
- Распознавание адресов вне Российской Федерации до уровня города.
- Одна улица может находиться в двух-трёх районах города, поэтому внутригородские районы не всегда определяем однозначно.
- Если района города нет в ФИАС, то он не влияет на результат разбора: например, оба адреса «Москва, Хамовники р-н, ул Остоженка» «Москва, Останкинский р-н, ул Остоженка» разберутся в «Москва, ул Остоженка» со статусом «гарантированный».

Используемые материалы

- ФИАС ФНС. Можно обновлять без перезапуска Фактора.
- Росстат (население в городах и населённых пунктах).
- Если у вас есть собственная полная база адресов, Фактор может использовать её вместо ФИАС.
- Справочник эталонных индексов от Почты России. Можно обновлять без перезапуска Фактора.

Площади и стоимости квартир

По адресу определяем:

- Площадь квартиры и частного дома по данным Росреестра.
- Среднюю стоимости квадратного метра в доме (по объявлениям в Интернете).

Ограничения

- Площади квартир и частных домов были выложены в свободный доступ Росреестром в 2013 году, после не обновлялись, поэтому новые данные не добавляются. Из этого источника мы получили порядка 60% квартир городов-миллионников и 23% квартир для остальных территорий РФ.
- Справочник стоимости квадратного метра в доме мы собираем по объявлениям о продаже в Интернете в нескольких источниках, обновляем раз в год.

Используемые материалы

ЕГРП Росреестра (площади квартир). Можно обновлять без перезапуска Фактора.

ДРУГИЕ ТИПЫ ДАННЫХ

Геокодирование

- Конвертируем адреса в GPS-координаты с указанием точности.
- Находим ближайшие к адресу клиента филиалы.
- Вычисляем положение объекта относительно МКАД и КАД: внутри или снаружи.
- Определяем ближайшие станции метро. Сделано для всех городов России, где есть метро: Москвы, Санкт-Петербурга, Казани, Самары, Новосибирска, Екатеринбурга, Нижнего Новгорода.

Можно обновлять справочник геокоординат без перезапуска Фактора.

Ограничения

- Геокодируем только те адреса, которые есть в ФИАС.
- Покрытие адресов из ФИАС геокоординатами не полное.
- Положение относительно МКАД и КАД — только для адресов Москвы и области, Санкт-Петербурга и области.
- Справочник границ МКАД и КАД, а также справочник метро нельзя обновить нагорячую.
- Расстояние до метро определяем по прямой.

Компании

- Выделяем организационно-правовую форму (ОПФ).
- Удаляем «мусор» из названия компании (мусорные слова, ИНН, КПП, почтовые адреса).
- Формируем название компании-ключа для последующей идентификации дубликатов.
- Заменяем старые ОКОГУ на ОКОГУ, действующие с 1 января 2012 года.
- Проверяем:
  - ИНН с учетом ОПФ по контрольной сумме, формату и справочнику кодов налоговых органов (СОУН). Дополняем справочник СОУН недостающими значениями из ЕГРЮЛ.
  - КПП по формату и справочнику СОУН. Дополняем справочник СОУН недостающими значениями из ЕГРЮЛ.
  - ОГРН по контрольной сумме, формату и справочнику кодов субъектов РФ.
  - ОКПО по контрольной сумме.
  - ОКВЭД по формату и допустимым значениям классификатора.
  - ОКАТО по контрольной сумме и формату.
  - SWIFT по формату.
  - КИО по длине и сравнение с ИНН.
  - ОКОГУ по справочнику.

БИК по формату.
Синтаксис URL в веб-сайте компании.

Используемые материалы

- Коды СОУН от ФНС — для проверки ИНН и КПП.
- Справочники ОПФ для стандартизации наименований и наш справочник опечаток в ОПФ.
- Справочник ОКВЭД.
- Справочник ОКОГУ.

Поиск компании в ЕГРЮЛ и предпринимателя в ЕГРИП

Проверяем реквизиты юридических лиц и индивидуальных предпринимателей по ЕГРЮЛ и ЕГРИП.

При поиске по ИНН/ОГРН восстанавливаем атрибуты:

- Полное наименование из ЕГРЮЛ
- Краткое наименование из ЕГРЮЛ
- Полное наименование на английском языке
- КПП/ОГРН/ИНН
- Код ОКОПФ
- Юридический адрес для ЮЛ и город регистрации для ИП.

Используемые материалы

ЕГРЮЛ и ЕГРИП по данным от ФНС. Обновляем раз в месяц. Можно обновлять без перезапуска Фактора.

Телефоны

- Выделяем код города из телефона.
- Приводим компоненты телефона к единому формату.
- Проверяем и восстанавливаем телефонные коды на основании почтового адреса.
- Преобразуем региональные телефонные коды в федеральные.
- Определяем тип телефона по его коду и ключевым словам в строке («факс», «моб», «дом», «раб» и пр).
- Выделяем мобильные.
- Выявляем несуществующие (например, «11111111»).
- Определяем оператора с учётом базы данных перенесённых номеров.

- Определяем регион оператора для мобильных номеров, адрес до города для стационарных.
- Проверяем соответствие длины телефона системе нумерации населенного пункта.
- Разделяем множества телефонов, заданных одной строкой, например, «моб 916 1510679, дом 320-78-10, рабочий 80951128912 доб. 3342» с учетом их типа.
- Заменяем телефонные коды устаревшим номерам.
- Присваиваем коды качества каждому номеру.
- Проверяем на существование номера в справочнике номерной емкости Россвязи.
- Распознаём номера телефонов в формате: <телефон>!<код города>, например, 7914216!495.

Ограничения

- Телефоны вне СНГ не распознаём.
- Из мобильных стандартизируем только российские номера.
- Оператор и регион определяем только для российских номеров.
- Прямые мобильные не идентифицируются.

Используемые материалы

- Справочник номерной ёмкости (Россвязь).
- База данных перенесённых номеров.
- Ресурсы Ростелекома.
- Информация о перенумерациях на АТС с сайтов региональных провайдеров связи.

ФИО и пол

- Разбиваем по компонентам (фамилия, имя, отчество).
- Определяем пол по фамилии, имени и отчеству.
- Выявляем инициалы.
- Исправляем опечатки.
- Проводим трансграфику: одиночные латинские буквы меняем на такие же кириллические (например, Сергей—Сергей).
- Исправляем окончания фамилии или отчества, если пол однозначно определён по другим полям.
- Заменяем уменьшительно-ласкательные имена полными.
- Понимаем ФИО, написанные латиницей. Алгоритм учитывает ГОСТы и то, как на самом деле люди могут писать имена.
- Выявляем несуществующие имена (например, «Не знаю»).
- Проставляем коды качества каждой компоненте.
- Склоняем: из именительного падежа в родительный, дательный или творительный.

Ограничения

Исходные ФИО должны быть в именительном падеже.

Используемые материалы

Собственные выверенные справочники:

- Распространённых фамилий, имён, отчеств.
- Распространённых опечаток.
- Справочники для перевода в родительный, дательный, творительный падежи.

Паспорта

Возможности

- Проверяем
  - Формат номера и серии документов, удостоверяющих личность (ДУЛ).
  - Дату выдачи и окончания срока действия ДУЛ с учетом даты рождения.
  - Паспорта по перечню недействительных паспортов от МВД. Перечень можно обновлять без перезапуска Фактора.
- Восстановление типа ДУЛ по формату серии и номера.
- Проверка СНИЛС.

Даты

- Распознавание дат в различных форматах и приведение их к единому результирующему формату.
- Проверка корректности указания дня рождения и возраста.
- Исключение некорректных дат, появившихся в результате автозаполнения из форм (например, «01.01.1970») при проверке.

E-mail

- Проверяем синтаксис и длину домена второго уровня. Синтаксис по RFC 2822 и по правилам популярных почтовых сервисов, отличным от RFC.
- Проверяем существования домена первого уровня по справочнику.
- Исправляем частые опечатки по справочнику.
- Извлекаем несколько e-mail из одной строки.

- Проверяем по справочнику одноразовых email-ов.

Ограничения

Не проверяем существование email-адреса и почтового сервера

Используемые материалы

- Справочник доменов первого уровня.
- Справочник аккаунтов, имеющих ролевую принадлежность (sales, support, etc.)
- Распространённые опечатки в написании доменов второго уровня

Модели автомобилей

- Распознавание моделей, заданных в неструктурированном формате с учетом опечаток, избыточной или недостаточной информации, синонимов:

  Alfa Romeo 156 –> ALFA-ROMEO, 156

  Chery Amulet –> CHERY, AMULET/A15

  Chery Tiggo 2.0 –> CHERY, TIGGO/T11

  IKCO Samand –> IRAN-KHODRO, SAMAND

  Kia Cee'd ED –> KIA, CEED

- Стандартизация моделей и приведение их к справочнику моделей ОСАГО (обновления производятся в рамках обновлений Фактор).
- Варианты распознавания моделей для случаев негарантированного распознавания, предлагаемые оператору в режиме реального времени:

  BMV 3 –> BMW M3, BMW X3, BMW Z3

  Fiat Ducato –> FIAT DUCATO (КАТЕГОРИЯ B), FIAT DUCATO (КАТЕГОРИЯ C), FIAT DUCATO (КАТЕГОРИЯ D)

  Kia K-серии –> KIA K2500, KIA K2700

Используемые материалы

Классификатор автомобилей auto.ru

БАЗОВЫЕ ПРАВИЛА

Правила, входящие в базовую поставку Фактор. Помогают преобразовать данные для миграции.

- Объединение двух и более полей в одно.
- Условный оператор.
- Копирование.
- Удаление или маркировка значения, содержащего «мусор».
- Замена по справочнику.
- Определение часового пояса РФ.

Обратная транслитерация

- Распознавание адресов, написанных на английском языке и транслите.
- Распознавание ФИО, написанных на английском языке и транслите.

Ограничения

- Нет литературного перевода с английского языка на русский.
- Распознавание компаний не включено (только некоторые правовые формы).

Идентификация террористов

- Проверка физических лиц: система проверяет, что стандартизованное ФИО контактного лица не похоже ни на одно ФИО из справочника на 80% и более (сравнение идет по алгоритмам, аналогичным алгоритмам идентификации дубликатов).
- Проверка компаний: система проверяет, что стандартизованное название компании не похоже ни на одну компанию из справочника на 80% и более (сравнение идет по алгоритмам, аналогичным алгоритмам идентификации дубликатов).
- Результат работы фильтра: заключение, является ли компания/физическое лицо террористом, название компании/физического лица террориста из базы террористов, % соответствия, идентификатор записи в базе террористов.

> Сведения о функциональности модулей идентификации дубликатов и домохозяйств Фактора

- Идентификация дубликатов
  - Возможности

Модули идентификации дубликатов позволяют производить идентификацию записей, которые похожи друг на друга, на основании заданного набора полей (например, ФИО, серия и номер паспорта и дата рождения).

Идентификация дубликатов

Идентификация дубликатов в Факторе описывается в виде группы *сценариев*.

Сценарий — правило, позволяющее определить, что некоторые записи являются дубликатами по определенному условию с определенной степенью похожести.

Для определения степени похожести записей используется набор *компараторов*, сравнивающий отдельные поля записи.

Компаратор производит сравнение одного или нескольких полей одного типа. Например, даты рождения двух физических лиц.

Возможности

- Специальные компараторы для адресов и имен, с учетом специфики сравнения данных компонентов
- Компараторы для сравнения групп телефонов и адресов.
- Компараторы для сравнения строк на похожесть.
- Возможность добавлять и редактировать группы сценариев и используемые ими компараторы в Фактор без программирования или помощи специалистов со стороны HFLabs.
- Возможность написания собственных компараторов на Java и подключения их в Фактор.
- Высокая скорость идентификации дубликатов.
- Генерация ключей, позволяющая осуществить идентификацию дубликатов в реальном времени и на инкрементальной основе, не сравнивая каждый раз всю базу.

Ограничения

Для запуска идентификации дубликатов данные должны быть стандартизированы.

Проверка по черным спискам

Поиск данных клиента (ЮЛ или ФЛ) на вхождение в справочники:

- террористов;
- иностранных публичных должностных лиц;
- любых иных списков, например, недобросовестных заемщиков.

Возможности

- Возможность подключить любые справочники различных форматов как черные списки;
- Гибкие правила поиска по похожести (на основе идентификации дубликатов);
- Высокая скорость поиска по черным спискам:
  онлайн: 1/4 секунды при миллионном справочнике;
  пакетная обработка: 5 млн записей в час.

Ограничения

Справочники для поиска предоставляются Заказчиком.

СДЕЛАНО В 2017

- Сделано в 2017
  - Статистика
  - Проверяем e-mail по правилам их создания на mail.ru, gmail.com и других популярных сервисах
  - Фактор проверяет существование имени, когда переводит имена с латиницы
  - Поможем сэкономить на рассылке sms
  - Переобработайте адреса — станет больше хороших
  - Определяем ближайшие станции метро
  - Улучшения адресов
    - Улучшаем почтовый индекс
    - Внутригородские районы в адресах
    - Улучшили разбор неполных адресов
    - Добавили безопасный учёт опечаток
    - Понимаем коды регионов в адресах
    - Расклеиваем типы в адресах
  - Новые поля по адресу из ФИАС
    - Получаем код ИФНС для физлиц из ФИАС
    - Собираем классификационный код ФИАС
    - Получаем состояние дома из ФИАС
  - Стандартизация компаний
    - Дополняем проверки ИНН
    - Выводим подробный статус компании из ЕГРЮЛ и ЕГРИП
    - Учитываем больше адресов при поиске юрлиц в ЕГРЮЛ
    - Находим ИП даже с опечаткой в имени или после смены фамилии
  - Поддержка изменений в справочниках налоговой
    - Защитились от сбоев налоговой, или Новый формат выгрузок ЕГРЮЛ и ЕГРИП
    - Поддержка нового формата ФИАС
    - Изменения в ФИАСе в апреле
    - Обнаружили неверные индексы в старых ФИАС
    - Различаем дубли в ФИАС
    - Поддержали новую структуру ФИАС
    - КЛАДР превратится в тыкву в новогоднюю ночь
  - Для разработчиков
    - Упростили автоматизацию скачивания ФИАС для Фактора
    - Новые параметры сервиса поиска адреса
  - Для поддержки

Перешли на java 8
Упростили скачивание справочника ФИАС
Пополнили базу знаний о проблемах интеграции с Сибель
Добавили версию чёрных списков
Переделали логирование Фактора

Статистика

- Обновили 64 справочников, не считая ежедневных ЕГРЮЛ и ЕГРИП.
- Закрыли 301 запрос в поддержке Фактора.

Проверяем e-mail по правилам их создания на mail.ru, gmail.com и других популярных сервисах

Раньше Фактор проверял формат e-mail только на соответствие стандарту RFC 2822.

Многие почтовые сервисы дополняют стандартные ограничения своими правилами, поэтому не все e-mail правильно получали код качества *Корректный*.

Теперь отмечаем e-mail на ручную проверку, если он не соответствует правилам почтового сервиса.

Мы добавили:

1. Проверку e-mail по правилам популярных почтовых сервисов:
    - yandex.ru, ya.ru
    - mail.ru, list.ru, bk.ru, inbox.ru
    - hotmail.com
    - yahoo.com
    - gmail.com
    - icloud.com
2. Распознавание кириллических логинов для доменов:
    - письмо.рф
    - е-письмо.рф

Исходный e-mail	Код качества		Комментарий
	Был	Стал	

Иванов.И.И@письмо.рф	Некорректный	Корректный	письмо.рф разрешает кириллицу и точки
Ivanov_I_I@gmail.com	Корректный	Некорректный	gmail запрещает нижнее подчеркивание

Если вы уже проверяете e-mail в Факторе, то проверка заработает автоматически после обновления на 8.9.

Фактор проверяет существование имени, когда переводит имена с латиницы

Полностью переделали алгоритм перевода ФИО с латиницы на кириллицу.

Задача перевода в общем случае нерешаема, потому что одна и та же латинская буква может означать разные русские, например G — это Г или Ж, IA — это либо Я, либо ИА.

Одно и то же имя можно написать разными способами. Знакомьтесь, Юлия Юрьевна:

- Ulia Ur'evna
- Iuliia Iurevna
- Yulia Yurevna

Раньше мы сложные случаи добавляли в словарь перевода. Всех случаев в словаре не учесть, поэтому часто получалось неправильно.

Теперь мы одновременно переводим всевозможными способами и проверяем результат по справочникам русских фамилий, имён и отчеств. Если попали однозначно, результат отдаём. Если возможны несколько вариантов, то выбираем один и ставим специальный код качества.

Новый алгоритм учитывает 12 стандартов транслитерации, их комбинации и то, как на самом деле люди пишут свои имена.

Если вы уже использовали старый вариант перевода ФИО — обновим в рамках поддержки. Если не использовали, то подключим как доработку. Обращайтесь к сотруднику HFLabs.

Поможем сэкономить на рассылке sms

Наши заказчики рассылают SMS через посредников либо напрямую через операторов связи.

За услуги посредников надо доплатить, поэтому напрямую через операторов рассылать дешевле.

Определить оператора номера телефона можно по данным Россвязи. Но Россвязь не сообщает, когда человек сменил оператора с сохранением номера. Смена операторов отражается в распространяемой ЦНИИ Связи базе данных перенесённых номеров (БДПН).

Мы подключили БДПН и теперь Фактор знает обо всех 7 млн переносах.

Отдаём обновлённые данные в поле "оператор":

- Если вы его получаете, то правильный оператор определится сразу после установки Фактор 8.7.
- Если поле "оператор" ещё не используете, то подключим как доработку.

Также можем выводить признак, менял ли абонент оператора.

Справочник выкладываем каждый день, обновляйте без перезапуска Фактора.

Переобработайте адреса — станет больше хороших

В релизах 8.3, 8.4 и 8.6 доработали алгоритмы адресов для учёта часто встречающихся особенностей.

Чтобы стало больше гарантированных адресов, необходимо обработать адреса со статусами Not_validated_has_ambi и Not_validated_has_unparsed_parts.

На обработку обязательно отправлять исходные строки с адресами, а не текущий результат разбора.

Определяем ближайшие станции метро

Для курьерских служб мы добавили определение трёх ближайших станций метро.

> Исходный адрес: г Москва, Стремянный пер, д 38.
>
> Координаты, полученные Фактором на основе адреса: GEO_LAT = 55.7275357, GEO_LNG = 37.6270583.
>
> Информация о метро получаем по координатам, дополняем названием ветки:
>
> 0,2км до м. Серпуховская (Серпуховско-Тимирязевская)
> 0,3км до м. Добрынинская (Кольцевая)
> 0,7км до м. Павелецкая (Кольцевая)

Расстояние до метро определяем по прямой.

Сделано для всех городов России, где есть метро: Москвы, Санкт-Петербурга, Казани, Самары, Новосибирска, Екатеринбурга, Нижнего Новгорода.

Улучшения адресов

Улучшаем почтовый индекс

Индекс для почтовой рассылки нужен для точной доставки писем и посылок.
Когда индекс не удаётся определить по данным Почты России и ФИАСу, оставляем исходный индекс. Но исходный индекс может противоречить разобранному адресу, поэтому теперь, если регион индекса отличается от региона адреса, то исходный индекс не попадёт в результирующий адрес.

Значения кода качества индекса для рассылки изменились! Нужно переделать интеграцию, если используете список кодов качества индекса. Новые значения соответствуют реальным уровням ФИАСа: сравнение старых и новых значений.

Внутригородские районы в адресах

Учли, что в адресе люди могут написать реальный район города, которого пока нет в ФИАС. Теперь даже если исходно указан район города не из ФИАС, он не помешает разбору.

Адрес *Санкт-Петербург, Центральный микрорайон, ул Мира, 12* станет *г Санкт-Петербург, ул Мира, д 12* со статусом *гарантированный*.

Улучшили разбор неполных адресов

Раньше если город пропущен и есть улица, адрес мог быть гарантированным только в одном случае: исходный индекс точно указывает на пропущенный город.

Теперь для адресов с пропущенным городом мы учитываем исходный индекс, даже если он неточный. Определяем регион по первым трём цифрам индекса и ищем варианты уже внутри региона. Если внутри региона такая улица одна, то адрес разберётся гарантированно. Если таких улиц несколько, Фактор отдаст несколько вариантов внутри одного региона.

Примеры гарантированных адресов, когда в регионе есть лишь одна такая улица:

- 197379 УЛИЦА ДОЛГООЗЁРНАЯ 12 197373, г Санкт-Петербург, ул Долгоозерная, д 12

Примеры неоднозначных адресов:

> 184209, КНЯЖЕГУБСКАЯ — в Мурманской области улица Княжегубская есть в двух населённых пунктах:

- Мурманская обл, Кандалакшский р-н, нп Зареченск, ул Княжегубская
- Мурманская обл, Кандалакшский р-н, пгт Зеленобор Княжегубская

108999 УЛИЦА БОГОРОДСКАЯ — Фактор предлагает 4 варианта разбора:

- Московская обл, г Ногинск, ул Богородская
- Московская обл, Щелковский р-н, деревня Назимиха, ул Богородская
- г Москва, г Троицк, ул Богородская
- г Москва, поселение Первомайское, деревня Пучково, ул Богородская

Не работает для популярных названий.

Добавили безопасный учёт опечаток

москва тулчанинов 119034, г Москва, пер Турчанинов

Понимаем коды регионов в адресах

Регион адреса хранится в виде кода во многих системах. Из них в Фактор на обработку приходят адреса с цифрами вместо слова.

Раньше, чтобы перевести код региона в наименование, требовалось подключить специальную настройку. Иначе адрес уходил на ручную проверку.

Теперь разбираем гарантированно из коробки:

- ,191028,78, УЛ. МОХОВАЯ,ДОМ 78 г Санкт-Петербург, ул Моховая, д 78
- 05 Мажалис Батырая дом 14 Респ Дагестан, село Маджалис, ул Батырая, д 14

Расклеиваем типы в адресах

Регулярно мы встречаем адреса, в которых тип склеен с наименованием: *гМосква, улМоховая*. Так происходит из-за ошибок при перекладывании из одной системы в другую.
Ошибку давно исправили, а слипшийся адрес остался в хранилище и пришёл на обработку в Фактор.

Раньше мы разбирали адрес неполностью и отправляли его на ручную проверку.

Сейчас разбираем полностью и гарантированно:

- МОСКВА УЛКРАСНОГОМАЯКА Д,10 г Москва, ул Красного Маяка, д 10
- респБашкортостан гИшимбай улБогданаХмельницкого д16 кв35 Респ Башкортостан, г Ишимбай, ул Богдана Хмельницкого, д 16, кв 35

Не будет работать, если одновременно допущена опечатка или тип передан в нестандартном виде.

Новые поля по адресу из ФИАС

Получаем код ИФНС для физлиц из ФИАС

Раньше по адресу получали из ФИАС только код ИФНС для юрлиц. Для отправки налоговой отчётности в правильную налоговую добавили ещё и ИФНС для физлиц.

Логика осталась прежней: если для объекта код есть, то возвращаем его, если нет — берём от "родителя".

Подключаем по запросу как доработку.

Собираем классификационный код ФИАС

Это цифровой код, похожий на код КЛАДР.

Нужен для:

- понимания адресообразования;
- интеграции с ведомствами.

Код выглядит так:

СС+РРР+ГГГ+ППП+СССС+УУУУ+ДДДД

- СС — код региона;
- РРР — код района;
- ГГГ — код города;
- ППП — код населенного пункта;
- СССС — код элемента планировочной структуры;
- УУУУ — код улицы;
- ДДДД — счетчик домов.

Например, *Московская обл, Ленинский р-н, с/п Развилковское, п Развилка, снт Поляна, ул Вишневая, д 34* 50014005002001300260001.

Подключаем по запросу как доработку.

Получаем состояние дома из ФИАС

Определяем состояние дома.

Состояние дома пригодится для понимания, является ли дом новостройкой или нежилым зданием. Например, в ФИАС 120 тысяч домов (0,5%) — новостройки.

Заполненность в ФИАС оставляет желать лучшего: 97,4% домов имеют статус без особого состояния.

Подключаем по запросу как доработку.

Стандартизация компаний

Дополняем проверки ИНН

Первые четыре цифры ИНН содержат код региона и номер налогового органа, проверяем их по перечню СОУН.
В ЕГРЮЛ и ЕГРИП есть ИНН, которые не проходят проверку по СОУН. В налоговой мы узнали, что это корректные ИНН и такие номера налоговых существуют.
Мы расширили справочник СОУН в Факторе дополнительными кодами СОУН: 9701, 9710, 9715, 9717, 9718, 9721, 9723, 9729.

ООО "СМАРТ ГРУПП", ИНН 9721048647

- Был код качества ИНН *NOT_VALID_TAX_CODE* — некорректный код СОУН.
- Стал код качества *GOOD* — хороший ИНН.

Добавили только массовые случаи. Например, с кодом 9721 в ЕГРЮЛ 4877 компаний.
В ЕГРЮЛ/ЕГРИП есть и ИНН, начинающиеся на другие «коды налоговых», которых нет в СОУН, например, 2040. Но их не добавили, потому что таких в ЕГРЮЛ по 1-2 компании и они все уже ликвидированы.

Выводим подробный статус компании из ЕГРЮЛ и ЕГРИП

Раньше Фактор отдавал только статус: действующая / в стадии ликвидации / ликвидирована.

Сейчас ставим дату ликвидации и в дополнение передаём подробный статус из ЕГРЮЛ и ЕГРИП.

Филиалам ставим статус головной компании.

Примеры:

МУНИЦИПАЛЬНОЕ ПРЕДПРИЯТИЕ "ДЕТАЛИ БЫТОВОЙ ТЕХНИКИ"

- ОГРН: 1086315007151
- Статус: Ликвидирована
- Дата ликвидации: 2008-07-25
- Подробный статус: "Прекращение деятельности юридического лица в связи с исключением из ЕГРЮЛ на основании п.2 ст.21.1 Федерального закона от 08.08.2001 №129-ФЗ".

Индивидуальный предприниматель ВЫСОТИНА ОКСАНА СЕРГЕЕВНА

- ОГРНИП: 304244710300031
- Дата прекращения деятельности: 2017-03-09
- Подробный статус: Индивидуальный предприниматель прекратил деятельность в связи с принятием им соответствующего решения

Учитываем больше адресов при поиске юрлиц в ЕГРЮЛ

Счастливым обладателям ЕГРЮЛ-фильтра.

Раньше при поиске учитывали только один адрес, а теперь сколько угодно. Это поможет найти больше компаний, потому что:

- в данных клиентов-юрлиц фактический и юридический адрес часто перепутаны местами.
- в ЕГРЮЛ у филиалов два адреса: местонахождения филиала и юридический адрес головной компании.

Чем больше адресов передано, тем ниже скорость: адреса стандартизируются перед поиском. Поэтому лучше ограничиться двумя или тремя адресами.

Настраиваем как доработку по запросу.

Находим ИП даже с опечаткой в имени или после смены фамилии

Раньше Фактор находил предпринимателя в ЕГРИП только если в дополнение к ИНН/ОГРН точно совпало ФИО.

Не все ИП находились по таким правилам, потому что:

- В реальных данных есть опечатки и инициалы.
- При смене фамилии в ЕГРИП — новая, а в базе контрагентов — старая.

Мы добавили сравнение ФИО по частям, теперь Фактор находит ИП при совпадении только фамилии или имени и отчества:

- ***Иванов**, ОГРН **315344300056295** ИП **Иванов** Евгений Александрович, ИНН 344309531858, ОГРН **315344300056295***

- *Сергеева **Антонина Петровна**, ИНН **790105092140** ИП Аниканова **Антонина Петровна,** ИН Н **790105092140,** ОГРН 309272115600013*

Настраиваем как доработку по запросу.

Поддержка изменений в справочниках налоговой

Защитились от сбоев налоговой, или Новый формат выгрузок ЕГРЮЛ и ЕГРИП

Зимой было два типа сбоев в сервисе налоговой службы для выгрузки данных из ЕГРЮЛ и ЕГРИП:

> Во второй половине января совсем не было обновлений для ЕГРЮЛ и ЕГРИП. По телефону нам объясняли, что неполадки связаны с проведением технических работ. Возможно, это была часть подготовки к переходу на новый формат выгрузки: с 1 апреля ФНС переходит на новый формат файлов с данными ЕГРЮЛ и ЕГРИП (приказ от 12 января 2017). На самом деле данные уже выгружаются в новом формате: по ЕГРЮЛ — с 31.01.2017, а ЕГРИП — с 10.03.2017.
>
> Несколько раз обновления дописывали в папку со вчерашними данными. Налоговая выкладывает изменения в директории с датами. Каждый день — новая директория. Например, для ЕГРИП: 21 февраля была информация 12 тысячах изменений, а 22 февраля — уже о 30 тысячах. Это значит, что, если не посмотреть в директории за прошедшие дни, то можно было потерять изменения по 18 000 индивидуальных предпринимателей.

Мы в Факторе поддержали новый формат и настроили полную синхронизацию с каталогами налоговой, чтобы получать даже те обновления, которые выложены задним числом.

Поддержка нового формата ФИАС

С 13 февраля налоговая изменила формат выгрузки ФИАС в виде dbf. Теперь вместо одного файла addrobj.dbf в архиве 86 файлов вида addrobj1.dbf, addrobj2.dbf … .

Мы поддержали изменение. ФИАС для Фактора можно скачать по ссылке и обновить по инструк ции. Неважно, стоит у вас 8.1 или 7.0 — везде будет работать.

Изменения в ФИАСе в апреле

В ФИАСе появился уровень зданий под номером 8, пока пустой. А ещё поменялись названия файлов, выкладываемых налоговой: ADDROBJ*ADDROB*
Поддержали эти изменения при генерации ФИАСа для Фактора.

Обнаружили неверные индексы в старых ФИАС

В апрельских ФИАС были массовые ошибки в почтовых индексах. Если используете ФИАС от 10 апреля, то лучше обновить на майский или июньский.

Или просто обновите Фактор на версию 8.4, в ней корректный ФИАС.

Различаем дубли в ФИАС

Для 3000+ населённых пунктов и улиц в ФИАС есть дубликаты: их названия, типы и родительские объекты полностью совпадают.
Проблема была в том, что Фактор возвращал неоднозначность и предлагал два одинаковых варианта — выглядит как полный бред, потому что при ручной обработке невозможно выбрать из двух одинаковых адресов.

Оказалось, что только в ~20% случаев в ФИАС задублирована одна и та же улица или деревня. Остальные записи только выглядят одинаково, но относятся к физически разным объектам.
Мы ещё научимся отдавать уточнения по физически разным объектам, чтобы было проще. Сейчас мы научились выделять точные дубликаты и перестали ставить им статус *неоднозначный*. Вот ключевые правила, по которым мы считаем адреса абсолютными дублями:

- Индексы и коды ОКАТО совпадают — выбираем любую запись.
- Оба индекса пустые, коды ОКАТО совпадают — выбираем любую запись.
- Один из индексов пустой, коды ОКАТО совпадают — выбираем запись с индексом:

    г Калининград, ул Заводская

    В ФИАСе две записи с одинаковым ОКАТО:

Адрес	Индекс	Код КЛАДР	Код ФИАС	ОКАТО
Россия, обл Калининградская, г Калининград, ул Заводская	236028	3900000100002 8000	4d7b1772-3d91-4b17-b8ee-c17e48ddf330	27401373 000
Россия, обл Калининградская, г Калининград, ул Заводская	-	3900000100012 0300	a1455f59-ca57-4104-ba11-e45b234f391d	27401373 000

    Выбираем запись с индексом, ставим статус проверки VALIDATED — гарантированный, все идентификаторы из ФИАС берём от верхней записи.

Если индексы либо коды ОКАТО **не** совпадают — перед нами 2 разных объекта.
Почему так? Отличающиеся ОКАТО означают, что улицы относятся к разным объектам территориального деления. Это одинаковые улицы в разных районах города, либо улицы в деревнях, которые полностью влились в город.

Подробнее на странице с новостями 8.3.

Поддержали новую структуру ФИАС

ФИАС — эталонный справочник адресов, в нём адреса хранятся в структурированном виде. Каждой части адреса соответствует свой уровень: регион, район, город, населённый пункт, улица, дом.

Рассмотрим пример: *188691 Ленинградская обл, д Кудрово, мкр Новый Оккервиль.*

В этом адресе есть:

- регион — *Ленинградская обл,*
- населённый пункт — *д Кудрово,*
- и микрорайон — *мкр Новый Оккервиль*. Стоп. У нас же нет уровня «микрорайон», есть «улица». Но какой-то необычный у неё тип — микрорайон?

В ФИАС *мкр Новый Оккервиль* относится к дополнительному уровню — «планировочной структуре».

На этом уровне располагаются дачные товарищества и микрорайоны. Сейчас 146 тысяч таких объектов.

Мы поддержали уровень планировочной структуры в Факторе, чтобы понимать объекты на этом уровне.

- Для адресов в формате КЛАДР, где не так много уровней, выводим планировочные структуры в населённый пункт или улицу — зависит от того, к какому объекту принадлежит.
  Пример: *188691 Ленинградская обл, д Кудрово, мкр Новый Оккервиль*. Мкр Новый Оккервиль — планировочная структура. Кудрово — деревня на уровне населённых пунктов. Выводим мкр Новый Оккервиль в поле для улицы.
- Если вы храните адреса в формате ФИАС, можем выводить новый уровень в отдельное поле.

Новый ФИАС не совместим со старыми версиями Фактора, потому что изменилась структура полей. Если самостоятельно обновляете ФИАС, то проверьте совместимость версий по инструкции.

КЛАДР превратится в тыкву в новогоднюю ночь

Налоговая опубликовала объявление.

С 1 января 2018 года КЛАДР исчезнет и останется только ФИАС. Фактор работает с ФИАС с 2015 года. Если ваши системы ещё используют КЛАДР, удалите его и используйте Фактор и Подсказки для приведения адресов к ФИАС. Наша статья о переходе с КЛАДР на ФИАС: https://habrahabr.ru/company/hflabs/blog/333736

Для разработчиков

Упростили автоматизацию скачивания ФИАС для Фактора

Стабилизировали пути, по которым надо скачивать ФИАС для Фактора.

В инструкции по обновлению справочника ФИАС см. примечание "Разработчкам"

Новые параметры сервиса поиска адреса

В Факторе есть способ вручную поискать адреса в ФИАСе. Он используется в РМАД и в Едином клиенте.

Также сервис поиска можно вызывать из других систем. Он поможет, если нужно найти город /населённый пункт по точному совпадению и без подчинённых объектов. В остальных случаях используйте Подсказки.

С версии 8.1 можно сократить число адресов в ответе:

- выводить только актуальные адреса,
- не выводить подчинённые объекты.

Новые параметры и примеры вызова сервиса поиска адреса — в документации.

Для поддержки

Перешли на java 8

Обновили требования к платформе. Для Фактора 8.8 и далее требуется java 8.

Упростили скачивание справочника ФИАС

Если вы самостоятельно обновляете ФИАС для Фактора, то теперь будет проще найти нужную версию.

Раньше для двух и более версий Фактора подходил один и тот же ФИАС. Теперь для каждой версии Фактора есть свой ФИАС.

Инструкция обновление справочника ФИАС стала проще.

Пополнили базу знаний о проблемах интеграции с Сибель

Проблемы при настройке коннектора

Добавили версию чёрных списков

На web-страницу с версиями справочников добавили дату каждого чёрного списка:

Версии справочников	
phone	2017-02-16
kladr	2017-02-16
geoCodes	2017-01-26
fias	2017-02-09
BlackList / terrorists.csv	2014-11-27
BlackList / Factiva.csv	2015-12-24

Переделали логирование Фактора

1. Появилась возможность записывать долгие запросы в отдельный файл. По умолчанию выключено.
2. Для расследования сложных проблем можно временно понизить уровень логирования без перезапуска Фактора.