Создание и ведение НСИ по нормализации адресов



Реализовано ЦНСИ-Компьютел для Главмосархитектуры

Программный комплекс
«BRADFORD»

www.computel.ru

ЗАДАЧИ

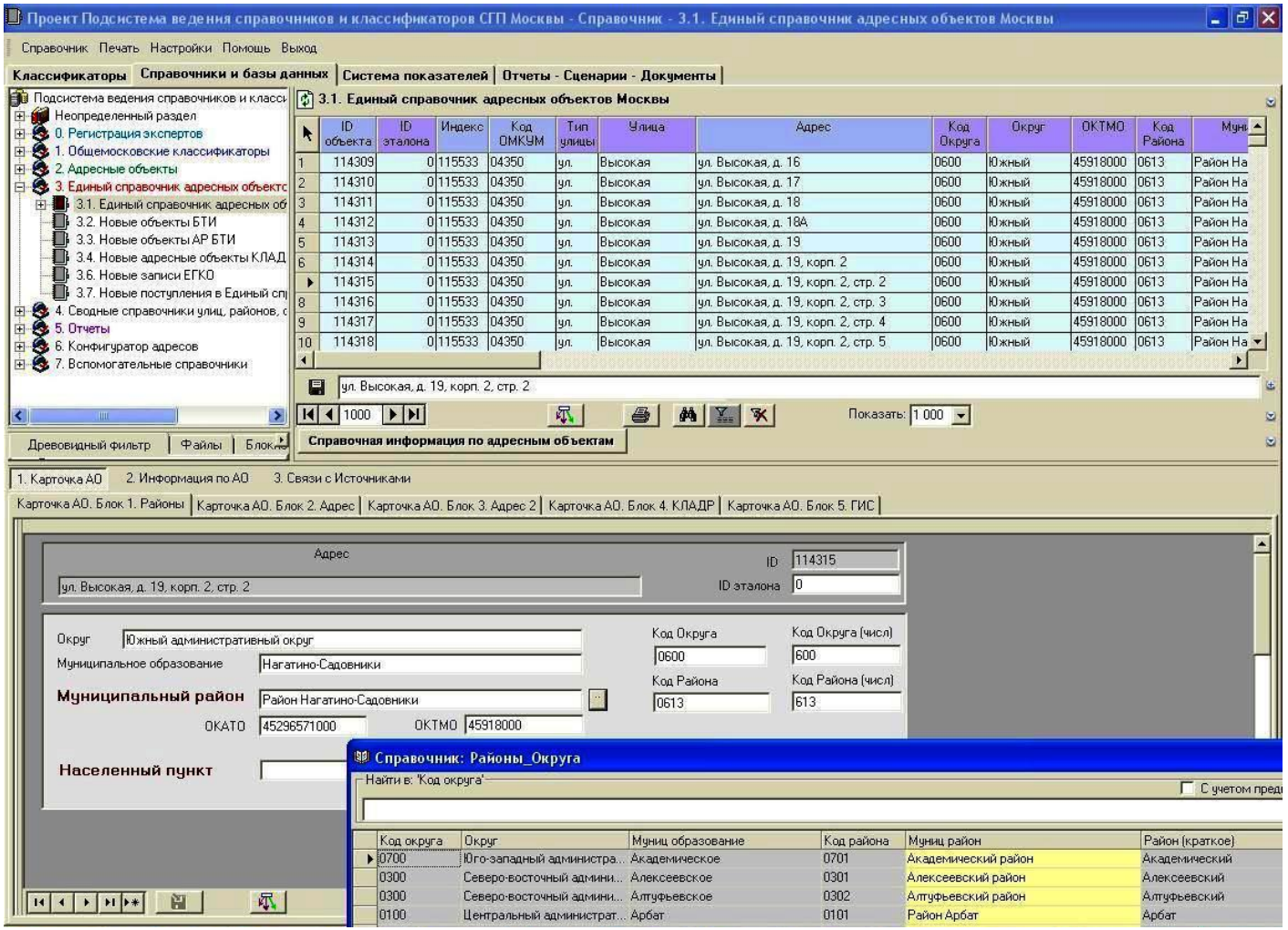
Идентификация адреса

Поиск дублей

Нормализация адреса

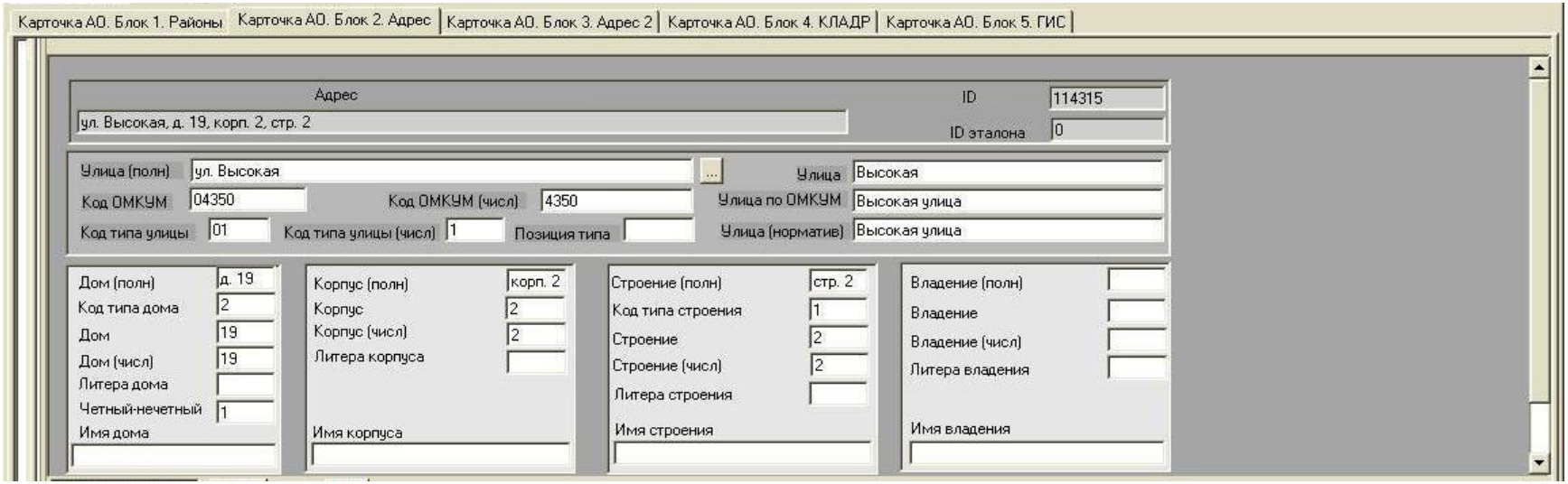
Технологии

Структура описания адресного объекта



Блок данных по округам, районам, населенным пунктам

Структура описания адресного объекта



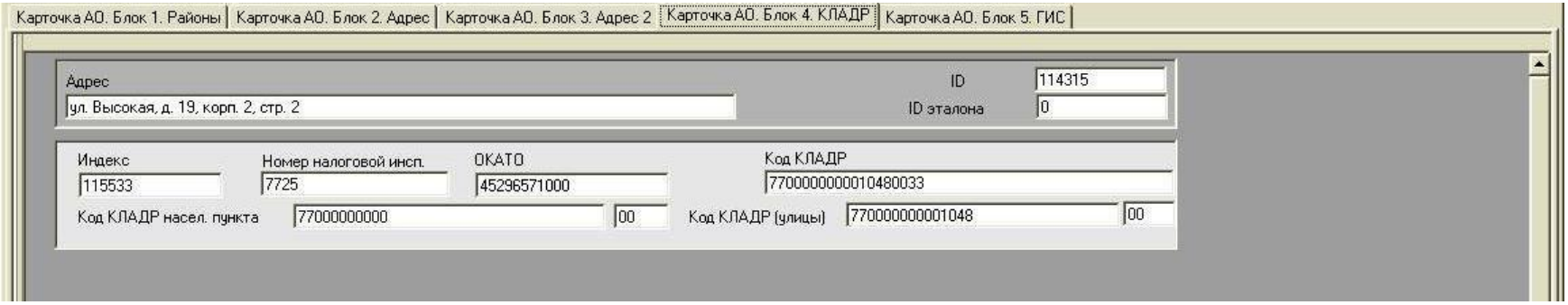
Блок данных основных структурных элементов адреса.

Структура описания адресного объекта



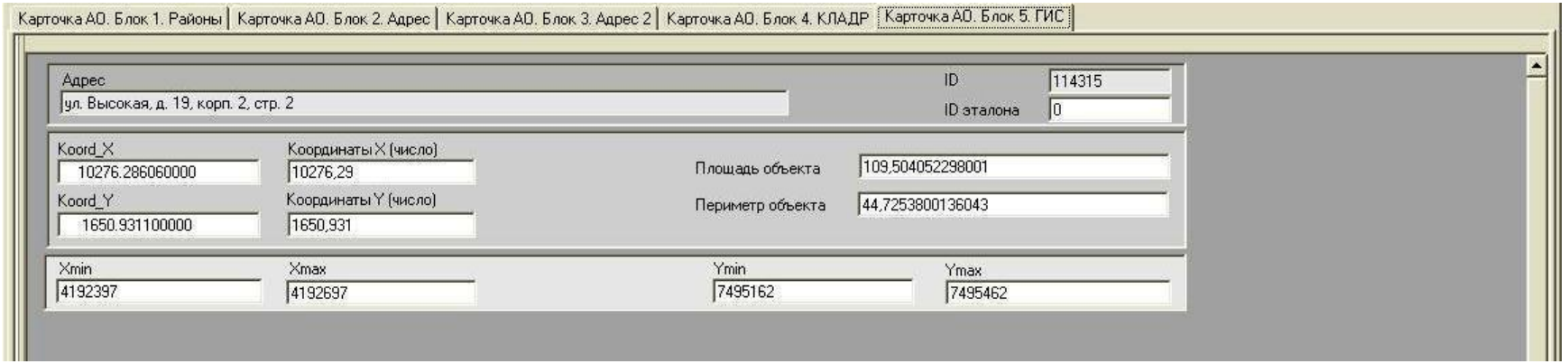
Блок данных второго адреса (для угловых зданий и сооружений).

Структура описания адресного объекта



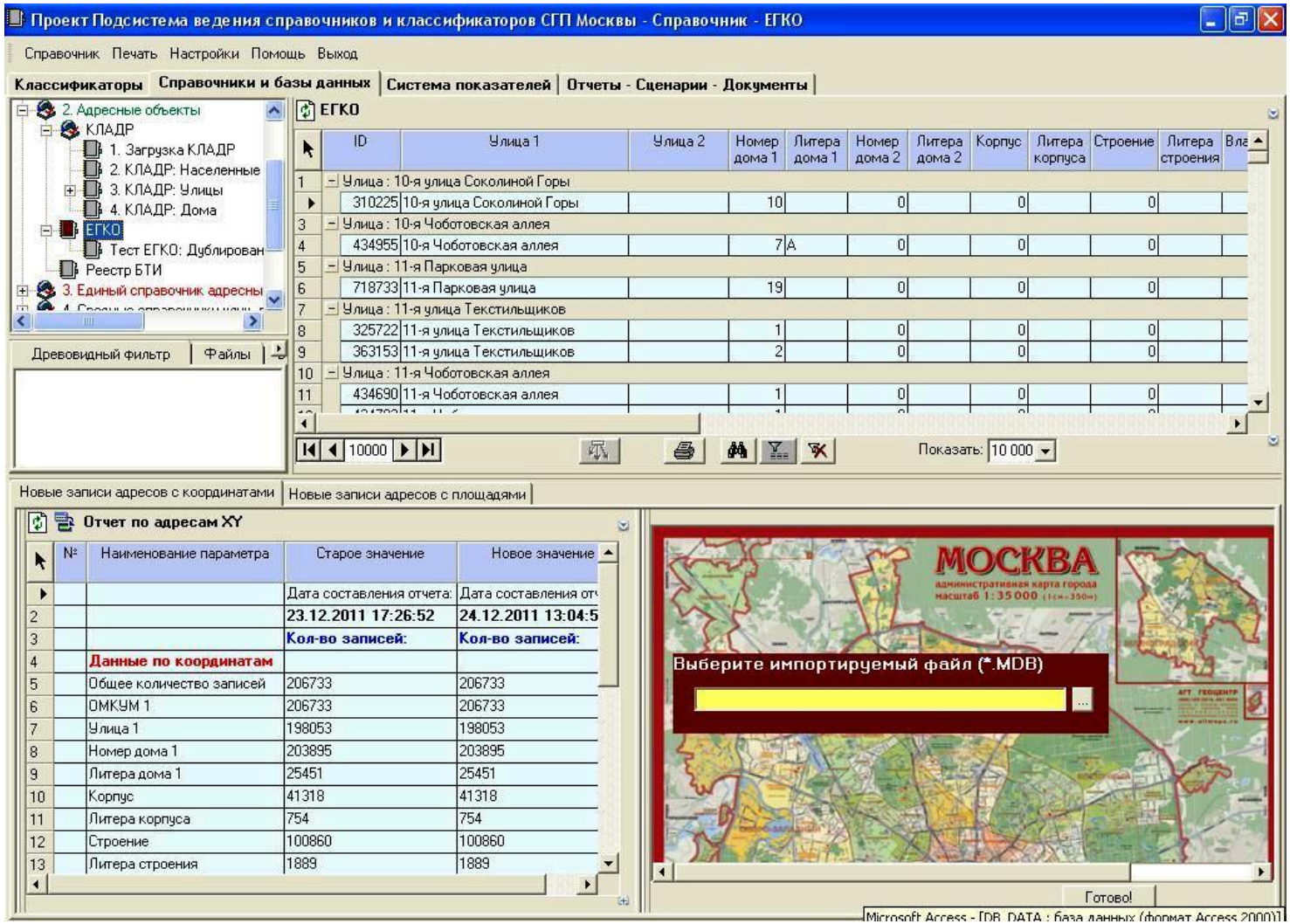
Блок данных по ключевым классификаторам (КЛАДР, ОКАТО и т.п.)

Структура описания адресного объекта



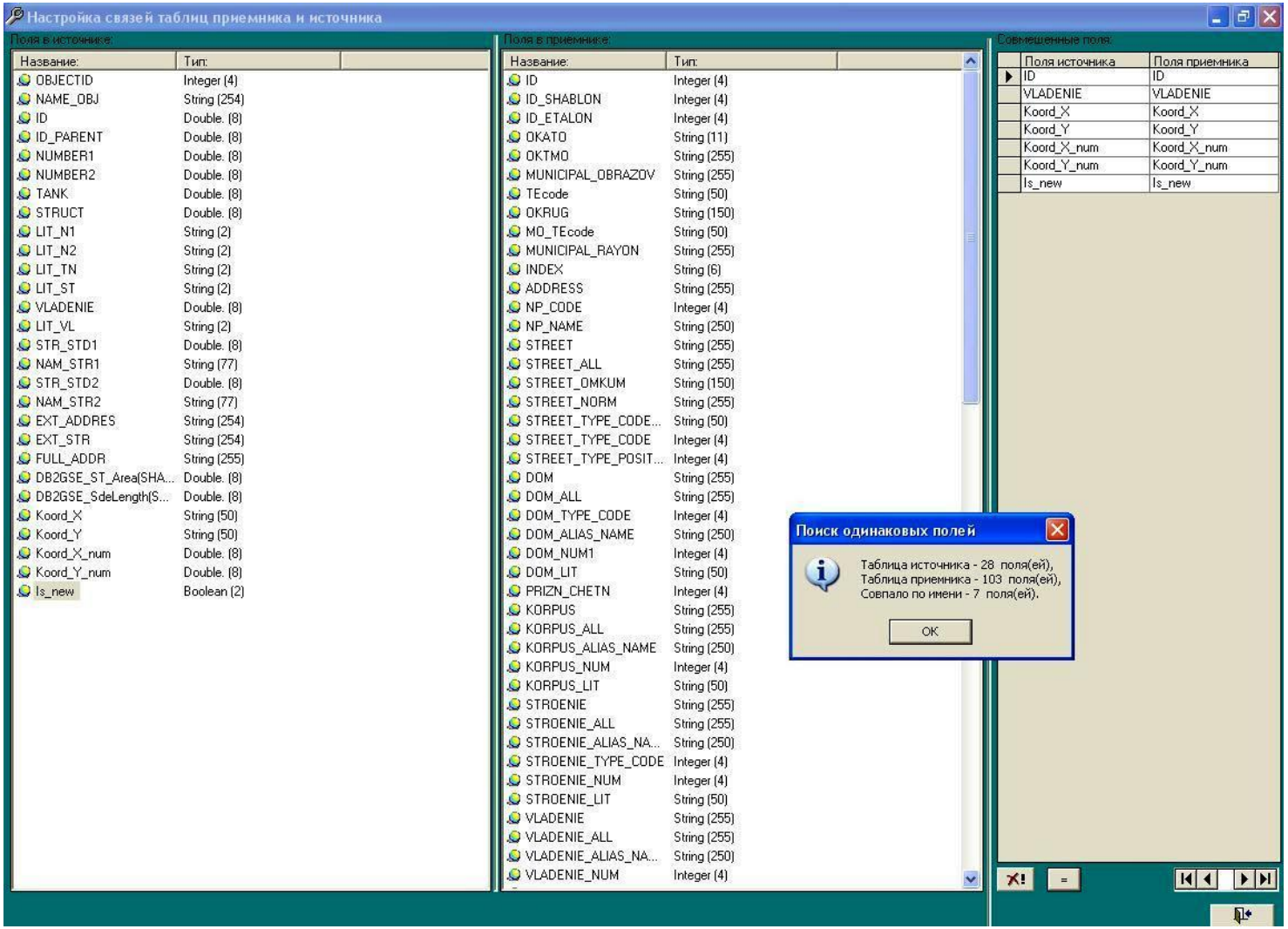
Блок геоинформации

Структурированный поиск по элементам адреса (импорт структурно организованных данных).



Вариант импорта структурно организованных данных (отдельные элементы адреса выделены в отдельные поля таблицы)

Структурированный поиск по элементам адреса (импорт структурно организованных данных).



Настройка связей между названиями полей элементов адреса импортируемой таблицы и приемной таблицы

Структурированный поиск по элементам адреса
(импорт структурно организованных данных).

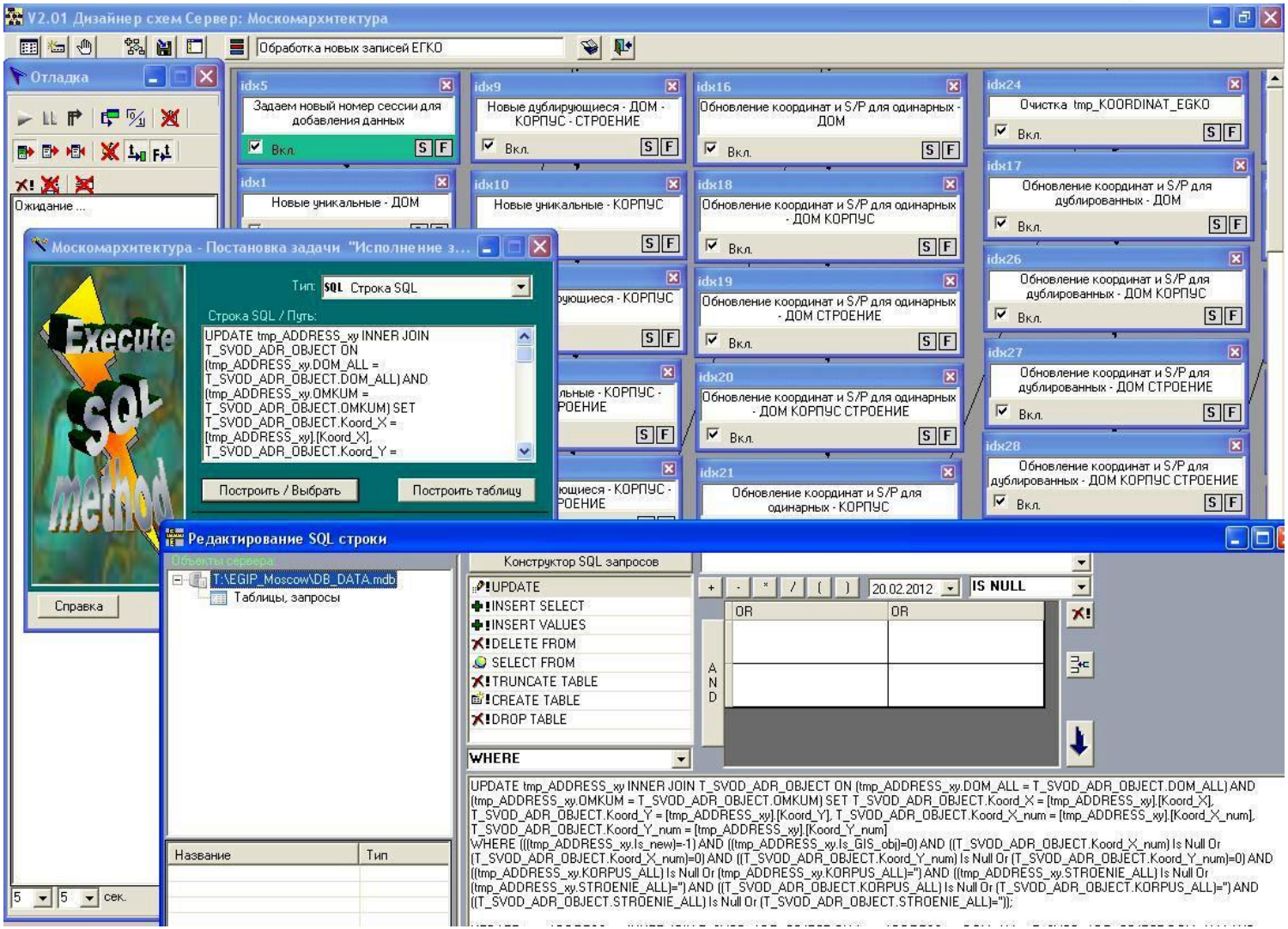
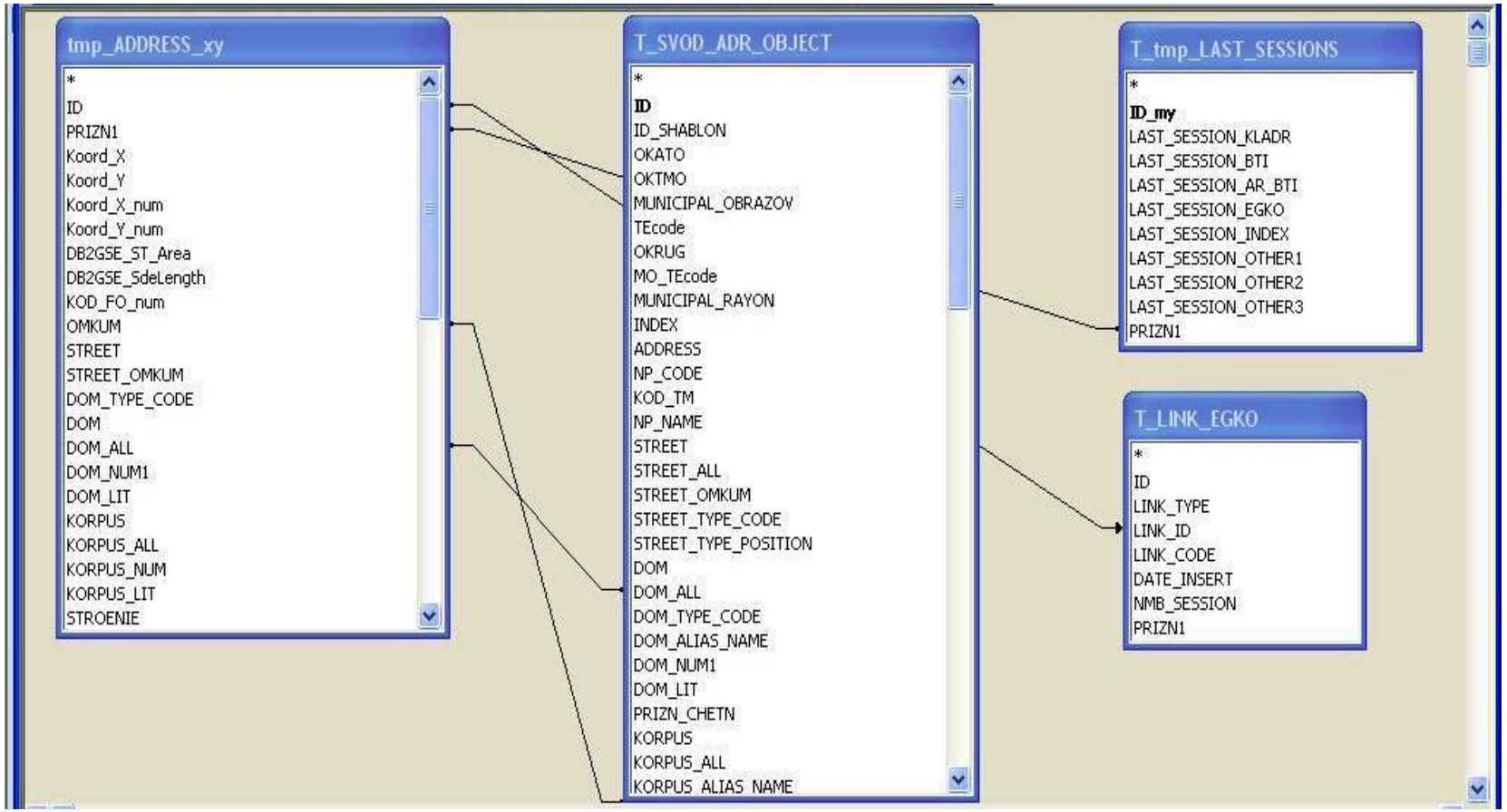


Схема последовательной идентификации адресного объекта в Сводном справочнике

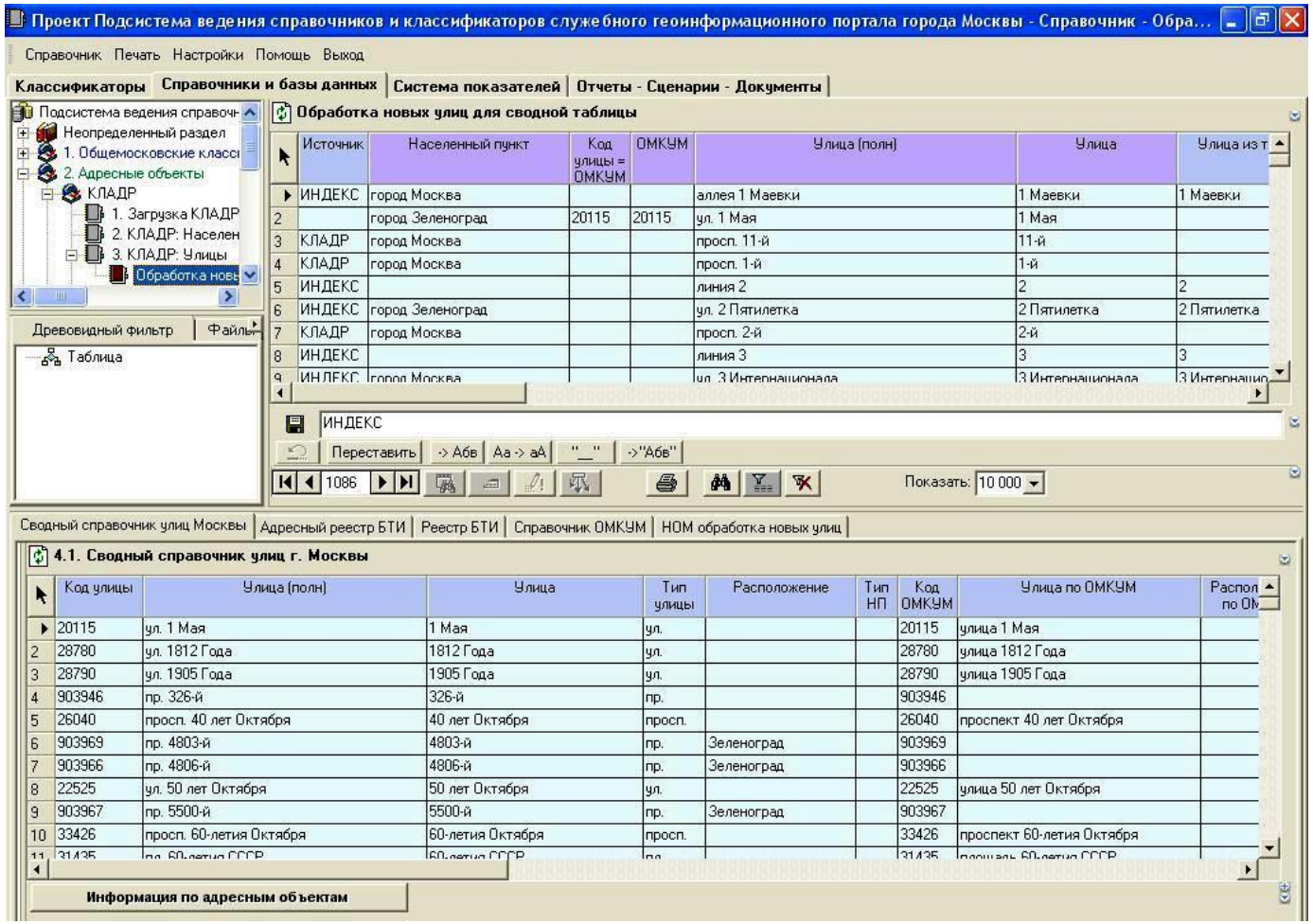
Структурированный поиск по элементам адреса
(импорт структурно организованных данных).



4. Объекты в виде SQL- запросов.

Каждый объект такой физической схемы обработки данных представляет собой один или несколько SQL-запросов или специализированных процедур ETL-сервера, например – идентификация по коду улицы и номеру дома

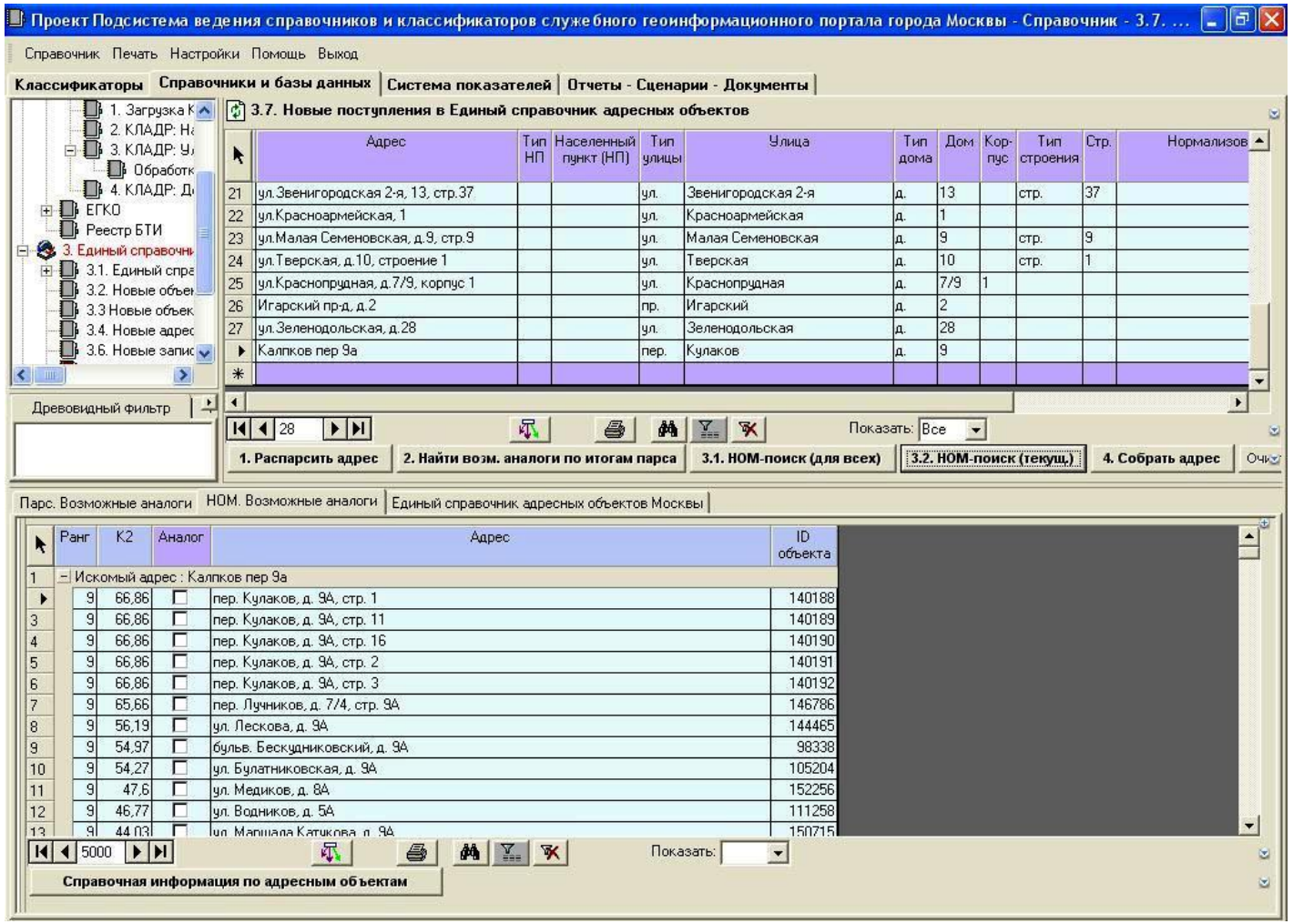
Структурированный поиск по элементам адреса
(импорт структурно организованных данных).



Процедура автоматической идентификации

Те записи, которые автоматически не идентифицируются, попадают в таблицу новых поступлений. Проверяется корректность присвоения этим записям кодов улиц, населенных пунктов и других основных параметров адреса, при необходимости производится правка этих кодов, после чего процедура автоматической идентификации запускается вновь.

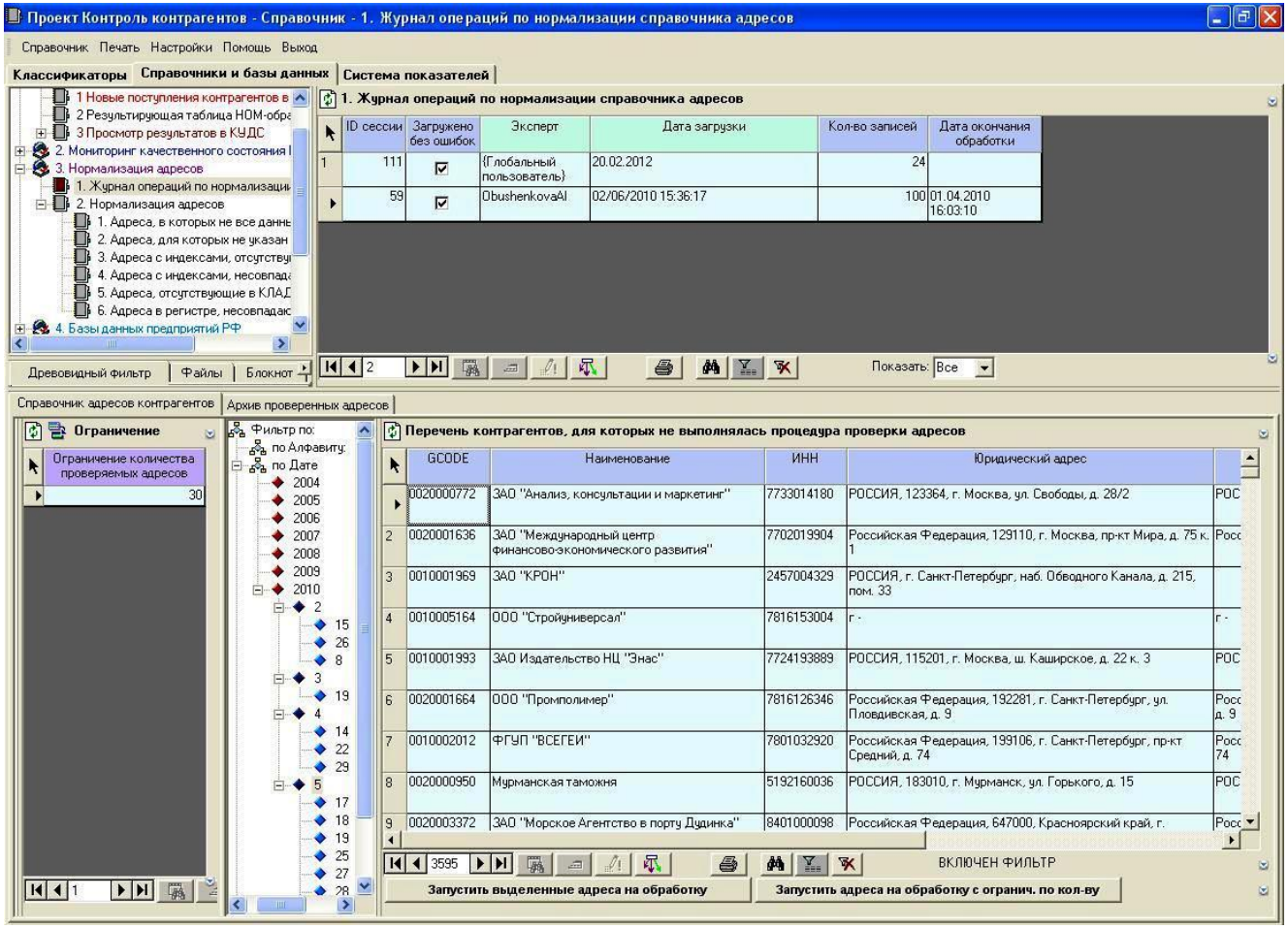
Структурированный поиск по элементам адреса
(импорт структурно организованных данных).



Специальный механизм поиска

Оставшиеся записи подлежат эвристическому разбору с помощью специального механизма поиска

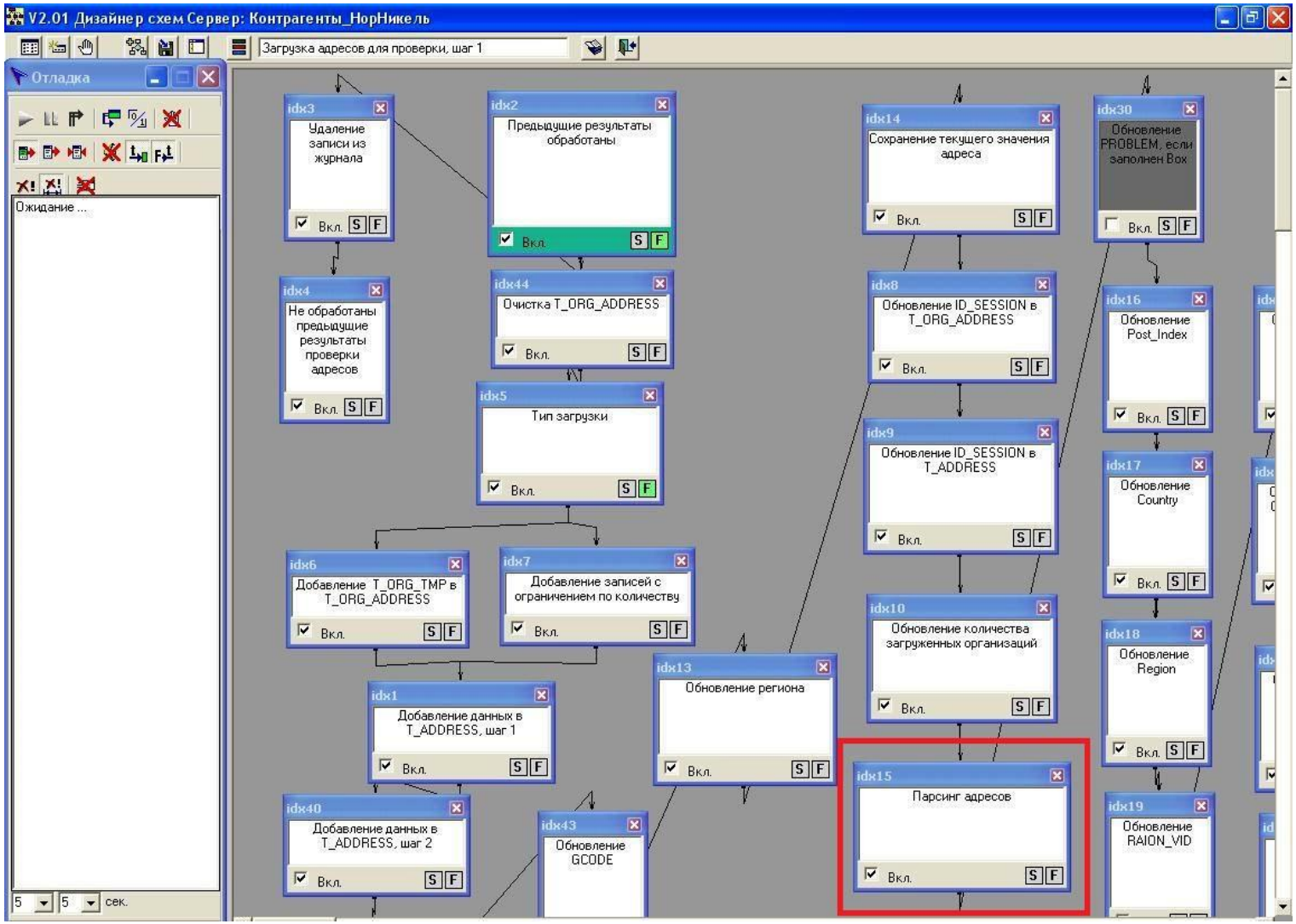
Структурированный поиск по элементам адреса
(импорт неструктурированных адресов).



Запуск поиска

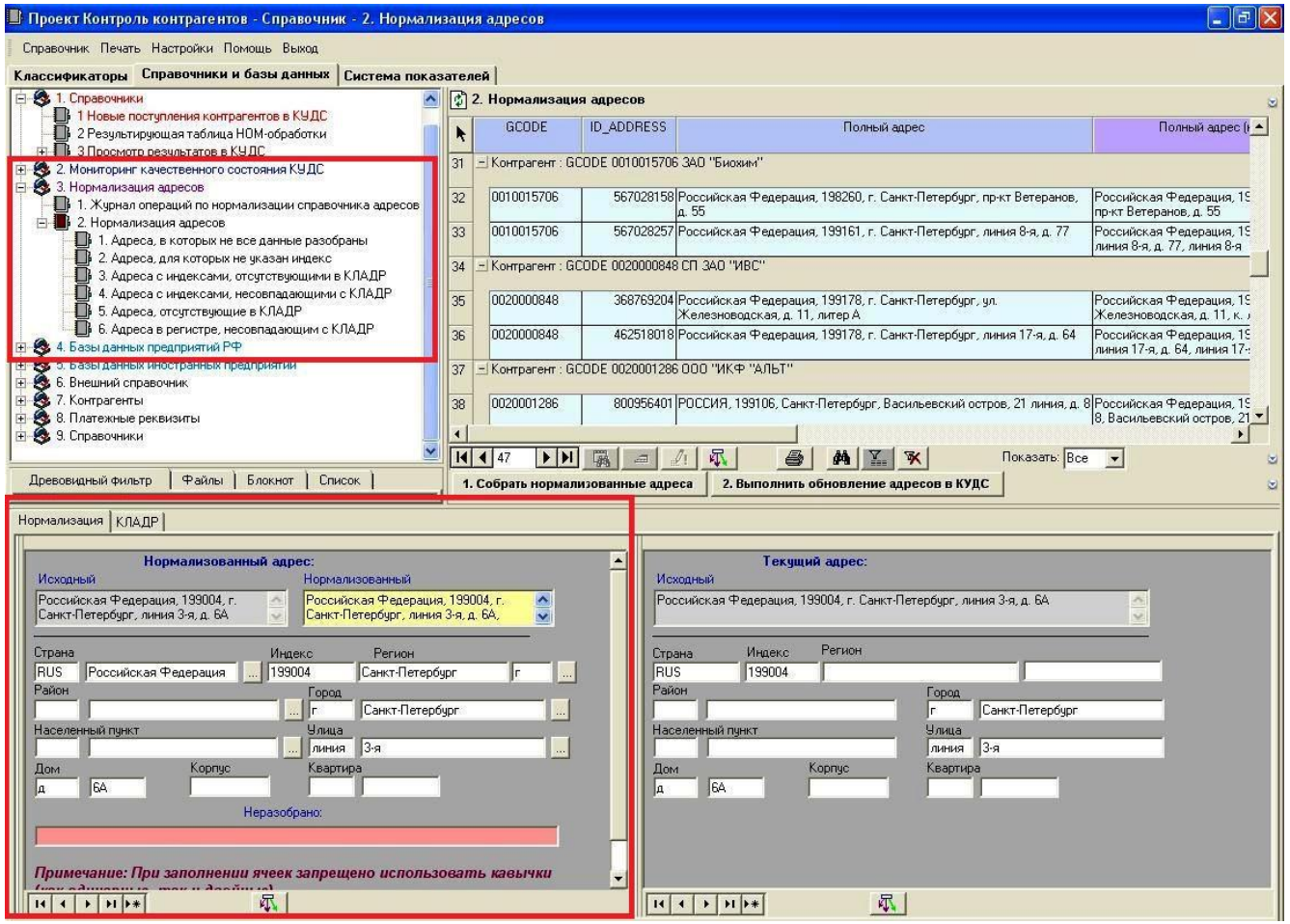
В случае, когда эксперты сталкиваются с необходимостью идентификации неструктурированного адреса, используется специализированная программа для разбивки (парсинга) единой текстовой записи на отдельные элементы адреса. При нажатии на кнопку «Запустить адреса на обработку» выполняется ряд операций, главной из которых является «Парсинг адресов»

Структурированный поиск по элементам адреса
(импорт неструктурированных адресов).



Парсинг адресов

Структурированный поиск по элементам адреса (импорт неструктурированных адресов).

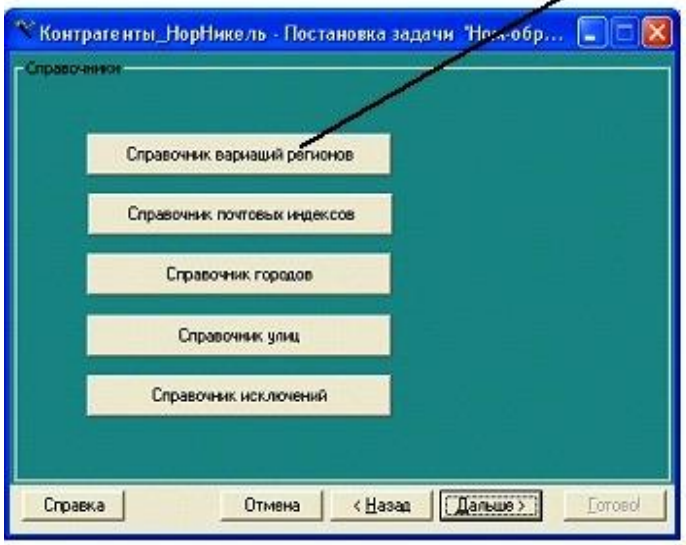


Подмассивы

В процессе выполнения данной схемы текстовый адрес разбивается на составные элементы, а весь массив анализируется и разделяется по типовым критериям на подмассивы (записи не разобраны, записи без индекса и т.п.) для удобства работы экспертов

Структурированный поиск по элементам адреса
(импорт неструктурированных адресов).

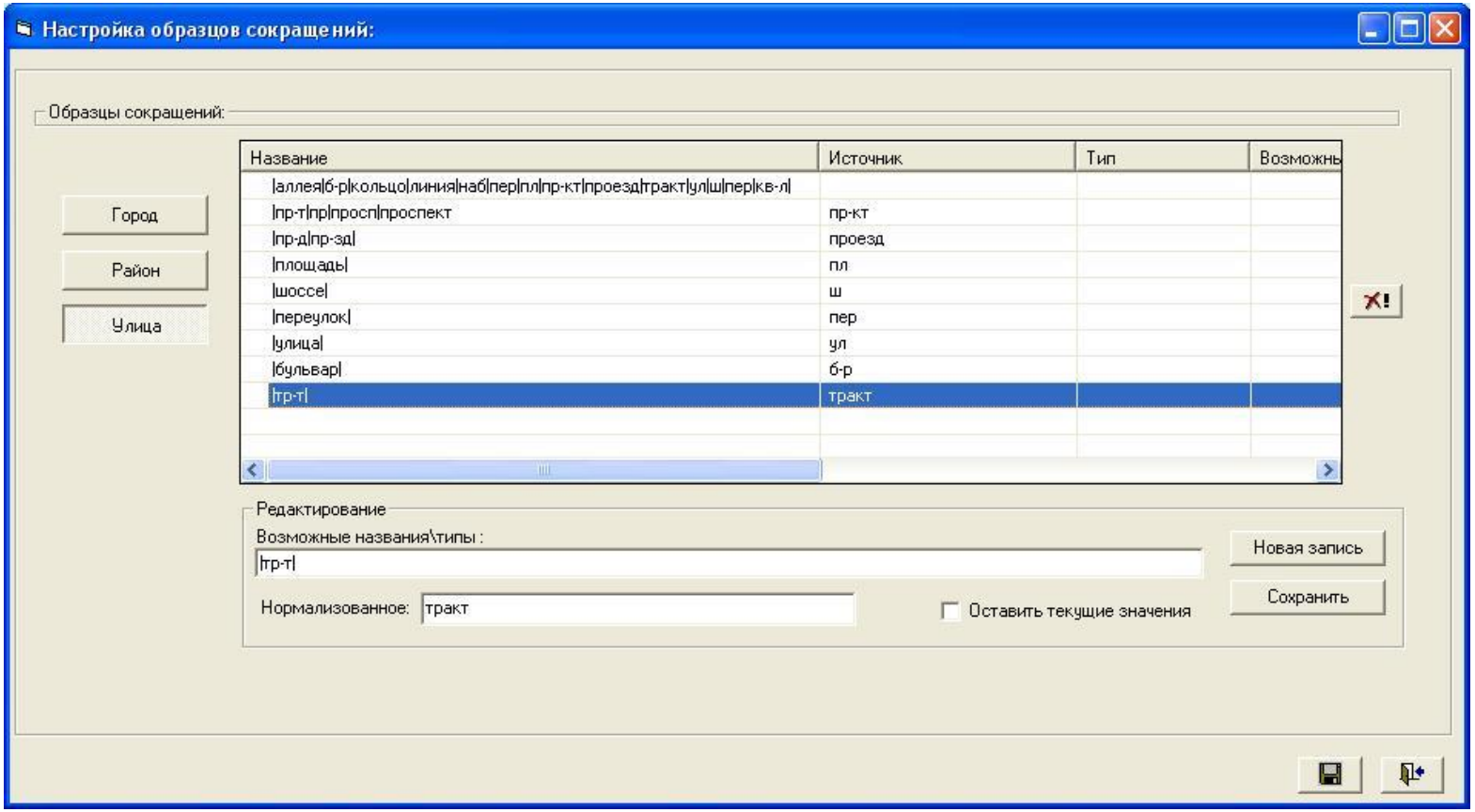
Для обеспечения правильности данной разбивки в программе предусмотрена возможность настройки серии вспомогательных справочников



01	Республика Адыгея
02	РеспюБашкортостан
02	Башкортастан
02	Республика Башкирия
02	БАШКОРТОСТАН
02	Республика Башкортостан
03	Республика Бурятия
04	Республика Алтай
05	Дагестан
05	Республика Дагестан
06	Республика Ингушетия
06	Ингушетия
07	КБР
07	Кабардино-Балкарская республика
08	Республика Калмыкия
09	КЧР
09	Карачаево-Черкесская республика
10	Республика Карелия
11	Республика Коми
12	РМЭ
12	РМЭ республика
12	Республика Марий-Эл
12	Марийская республика
12	Республика Марий Эл
13	Республика Мордовия
14	Республика Якутия
14	Республика Саха
15	РСО-Алания
15	Республика Алания
15	Республика Северная Осетия
16	Татарстан
16	Республика Татария
17	Республика Тыва
17	Республика Тува
18	Удмуртская республика

Справочник вариаций наименований регионов

Структурированный поиск по элементам адреса
(импорт неструктурированных адресов).



Справочник образцов сокращений

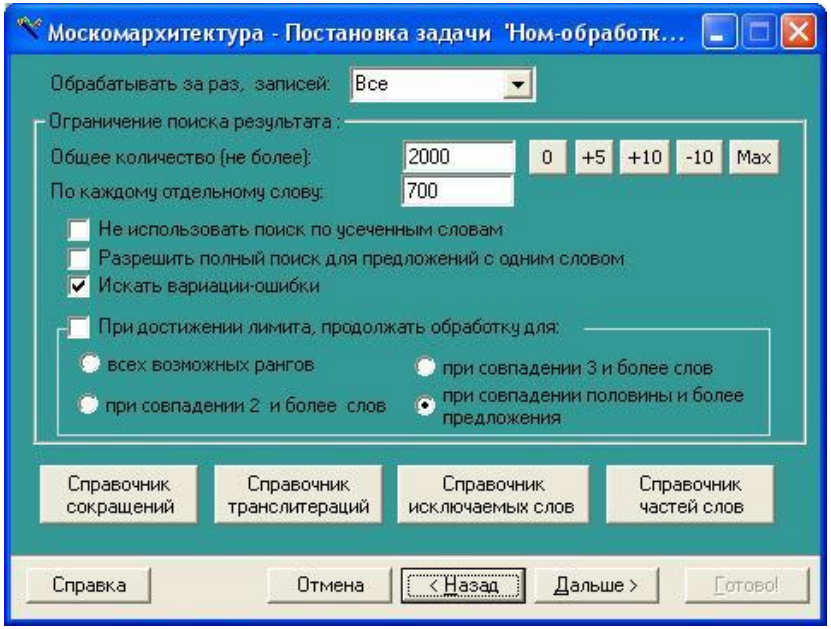
После структурирования адреса, дальнейшая работа по его идентификации (поиску соответствия, дублей) выполняется по порядку, описанному выше для структурированного поиска.

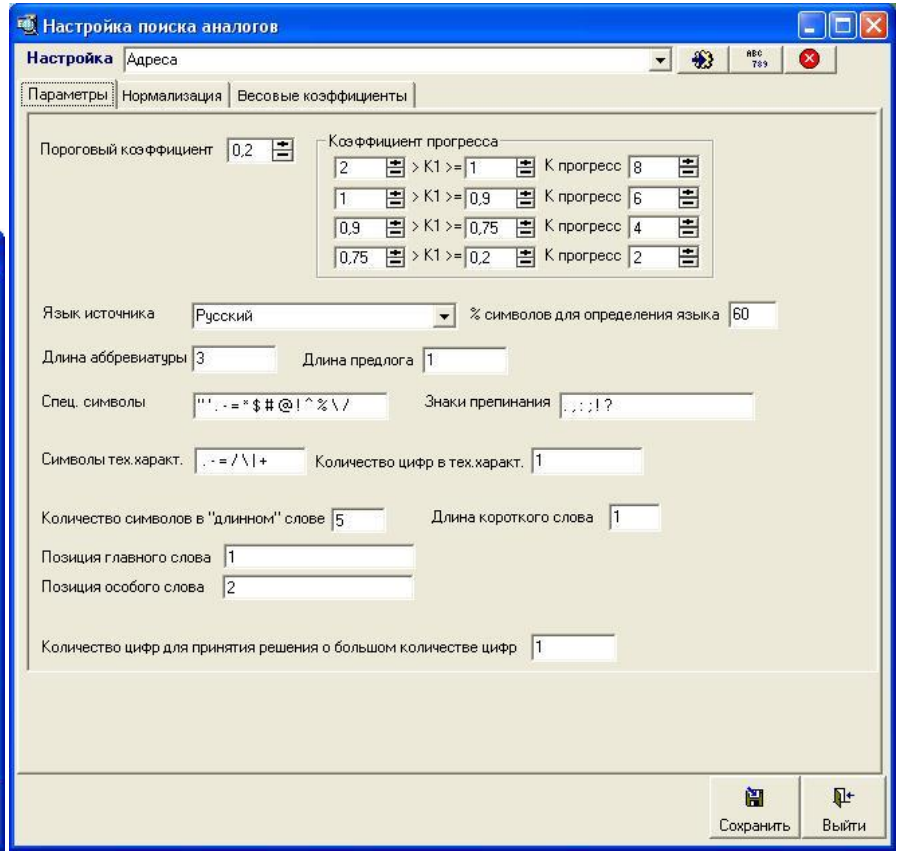
Эвристический поиск по полному адресу

- Для ручного поиска возможных аналогов в сводном массиве адресных объектов настроены несколько механизмов, в том числе, - механизм эвристического поиска с использованием алгоритмом нечеткой логики, с расчетом коэффициентов релевантности для найденных аналогов. Для усиления точности поиска созданы вспомогательные справочники – весовых коэффициентов слов, исключаемых слов, замен и т.п.

- Эксперту НСИ возможно потребуется использовать механизм тюнинговых настроек для усиления качества эвристического подбора аналогов. В настоящий момент выполненная настройка такого механизма уже в достаточно высокой степени учитывают специфику поиска именно по адресной базе данных.

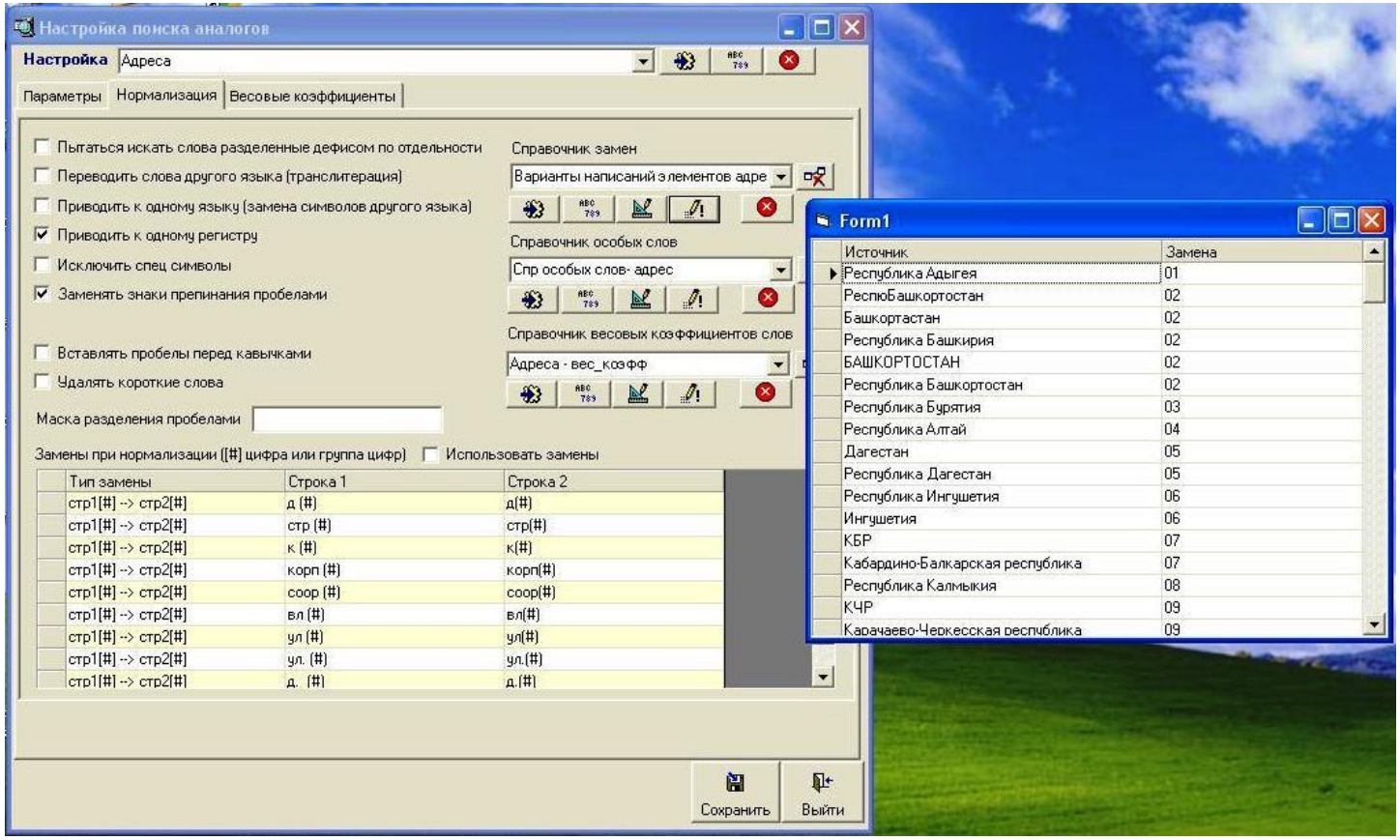
Эвристический поиск по полному адресу





Настройка механизма поиска

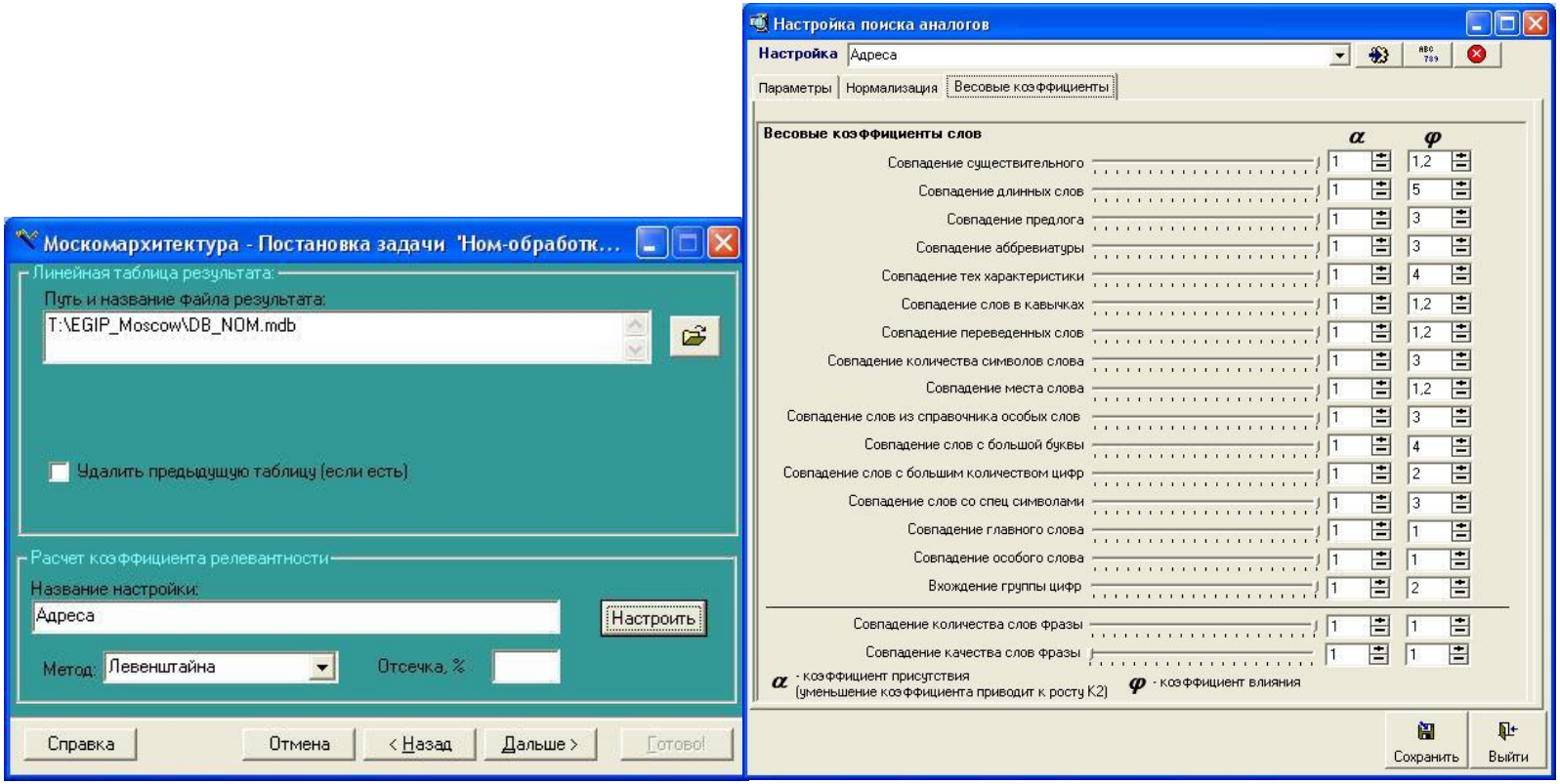
Эвристический поиск по полному адресу



Настройка механизма поиска

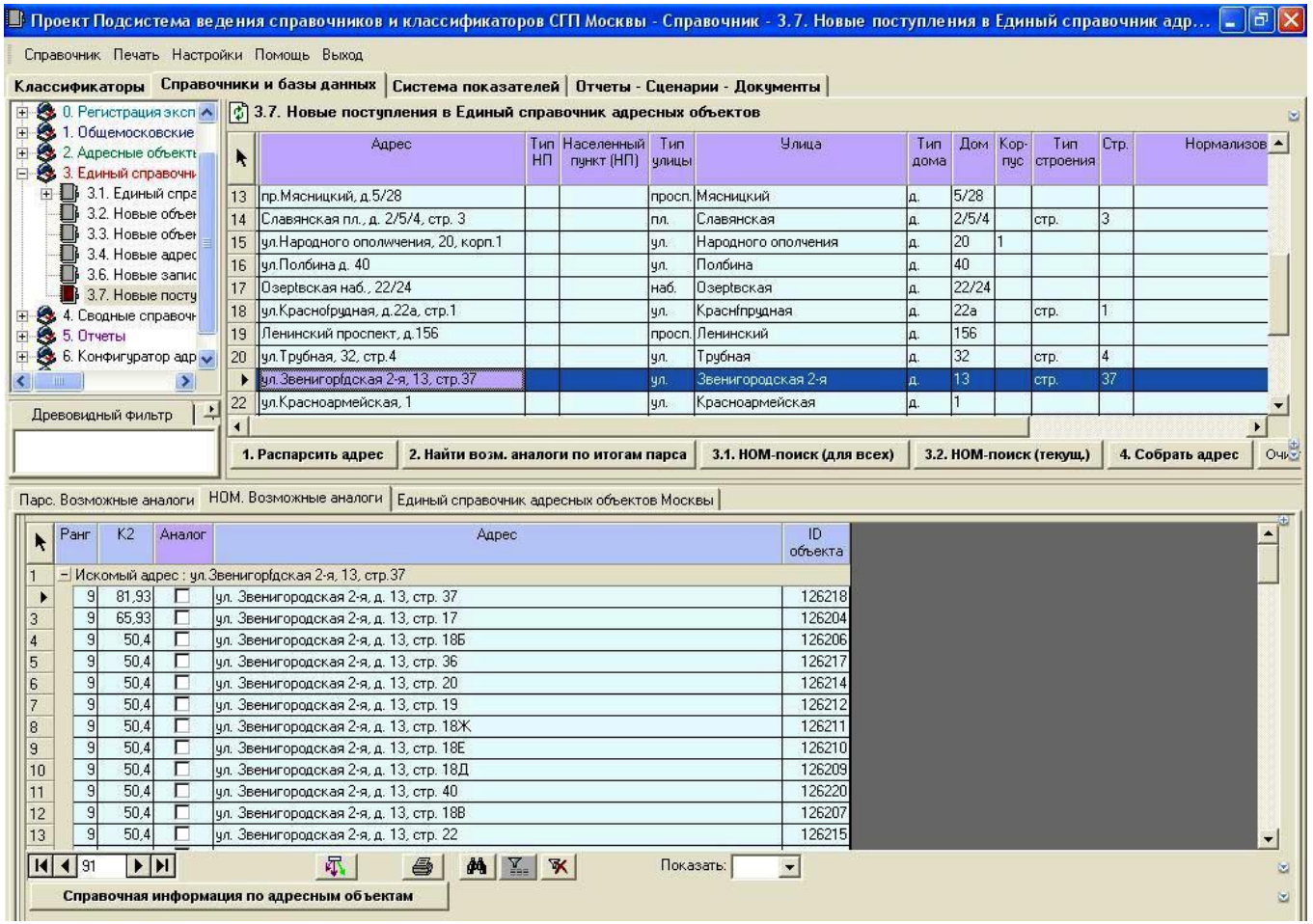
С помощью использования справочников замен, сокращений, транслитераций, исключаемых слов и т.п. возможно обеспечить адекватный поиск по устойчиво ошибочным вариантам названия, например, правильно разбирать «Екатеринбургскую область»:

Эвристический поиск по полному адресу



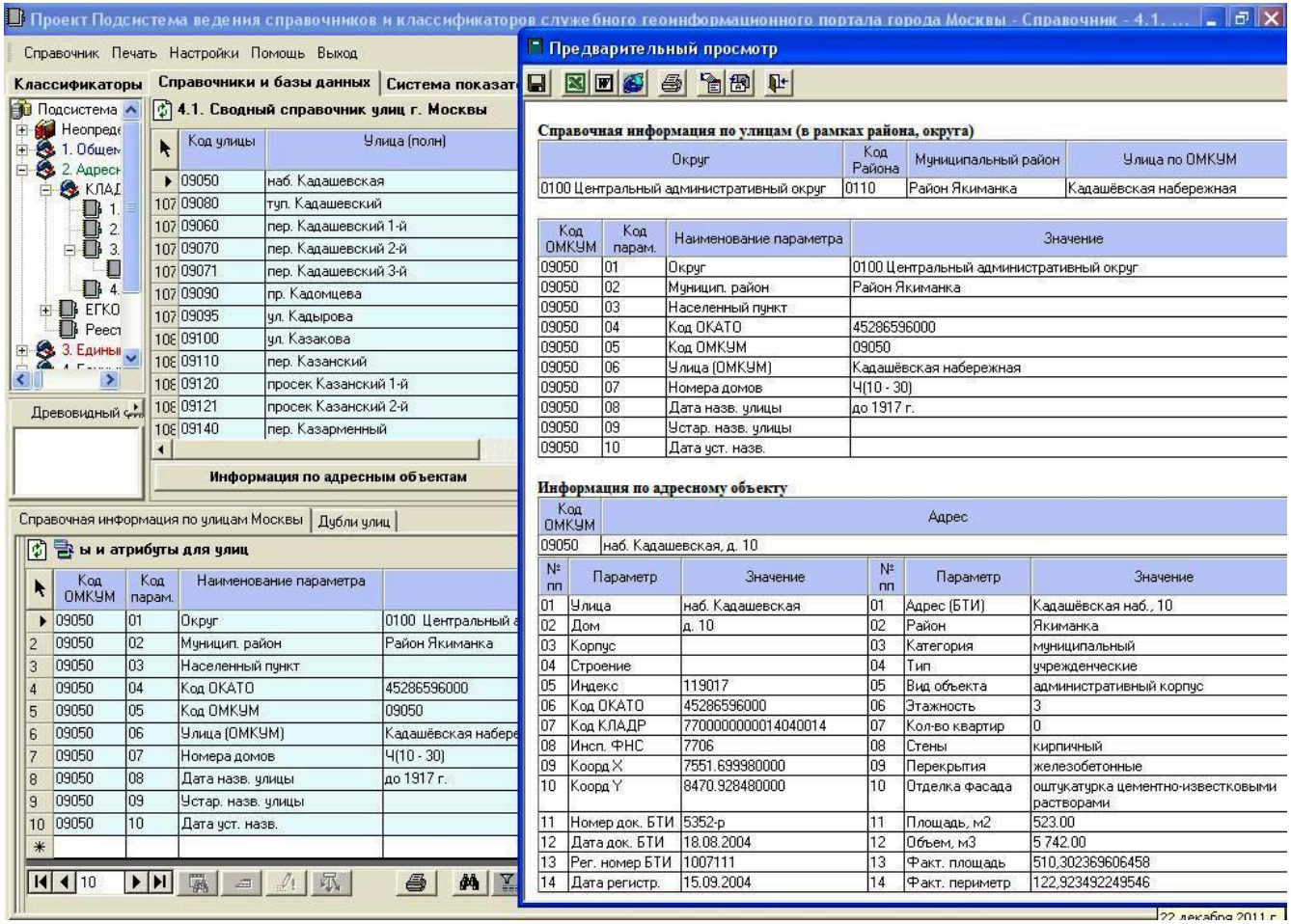
Примеры настройки механизма эвристического поиска

Эвристический поиск по полному адресу



Результаты работы такого механизма поиска (для примера в центре названия улицы специально две буквы были заменены на латиницу):

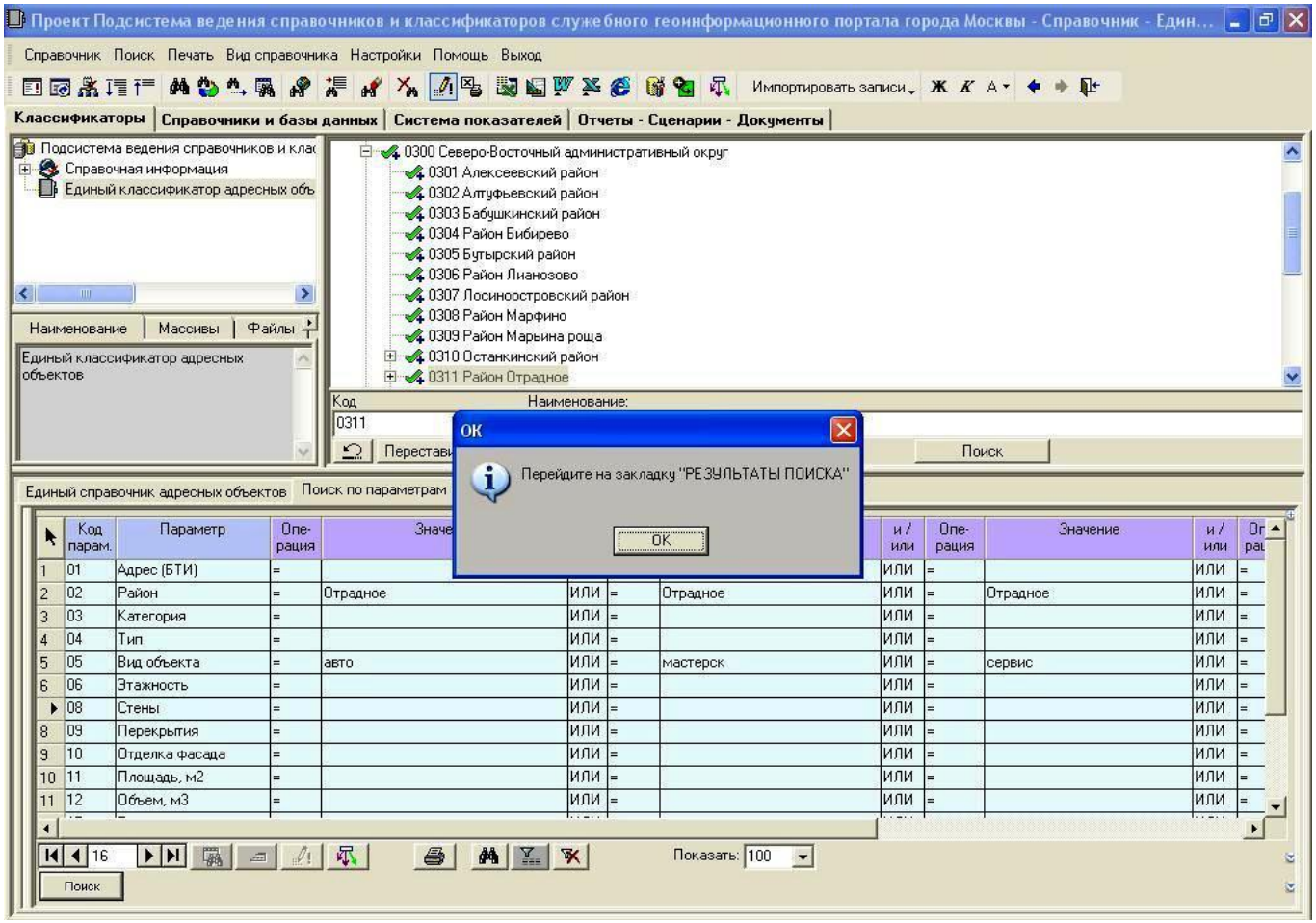
Поиск по дополнительным характеристикам адресных объектов



Пример дополнительной информации из БТИ

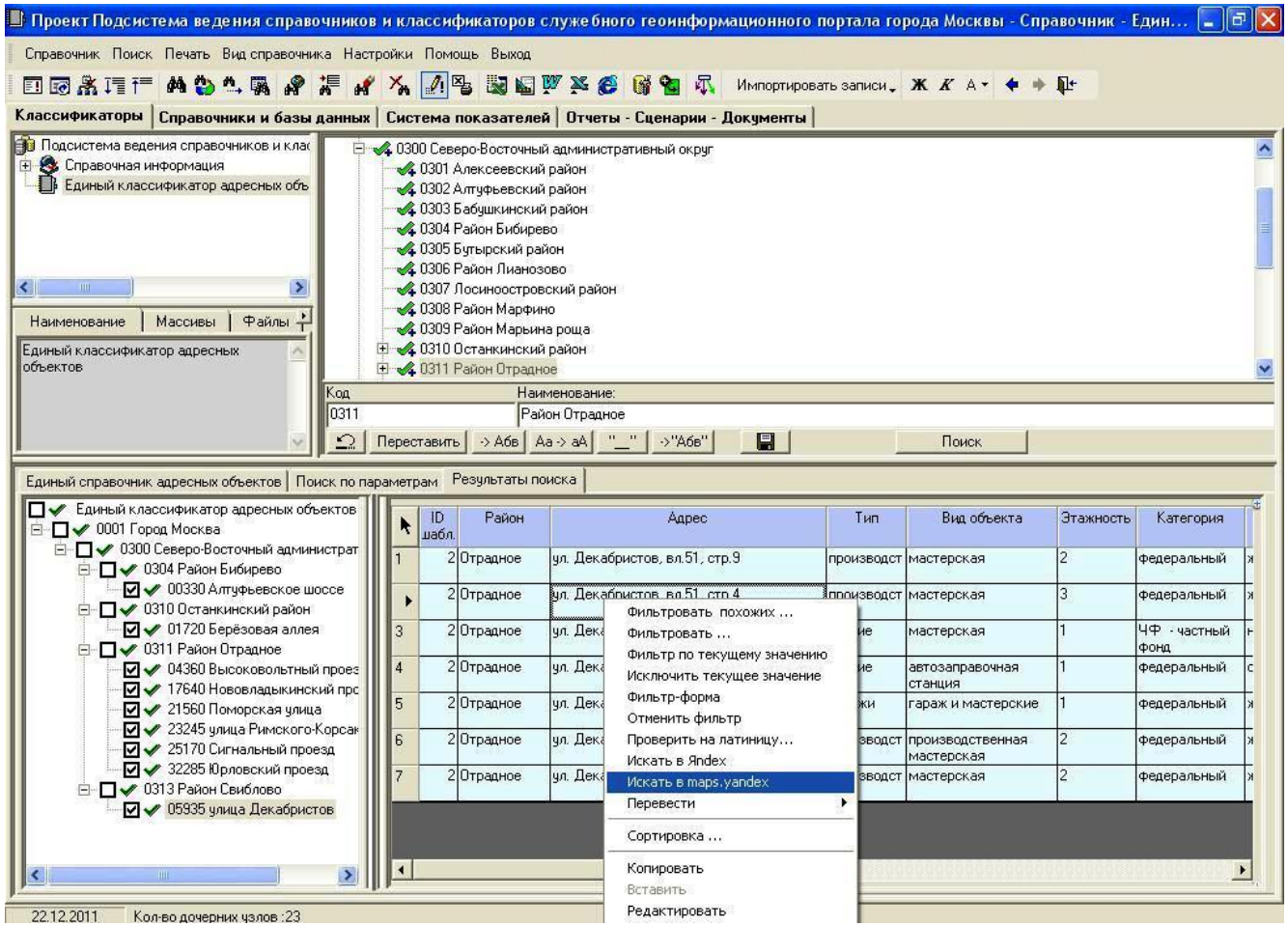
В случае, если в Сводный справочник адресов удается подключить дополнительную информацию (наименование организаций, фамилии адресатов, назначение адресных объектов и т.п.), то возможно организовать параметрический поиск по этим дополнительным параметрам

Поиск по дополнительным характеристикам адресных объектов



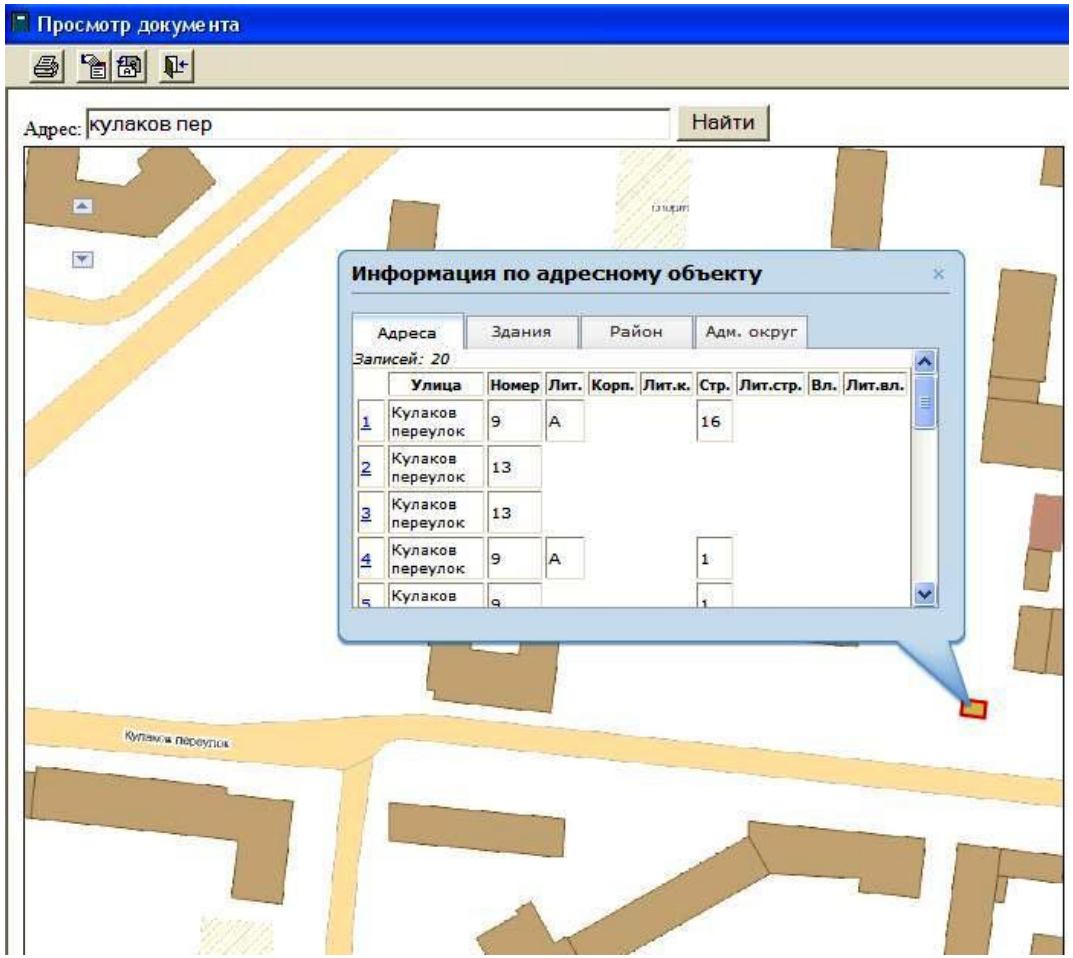
Окно задания условий параметрического поиска по дополнительной информации

Поиск по дополнительным характеристикам адресных объектов



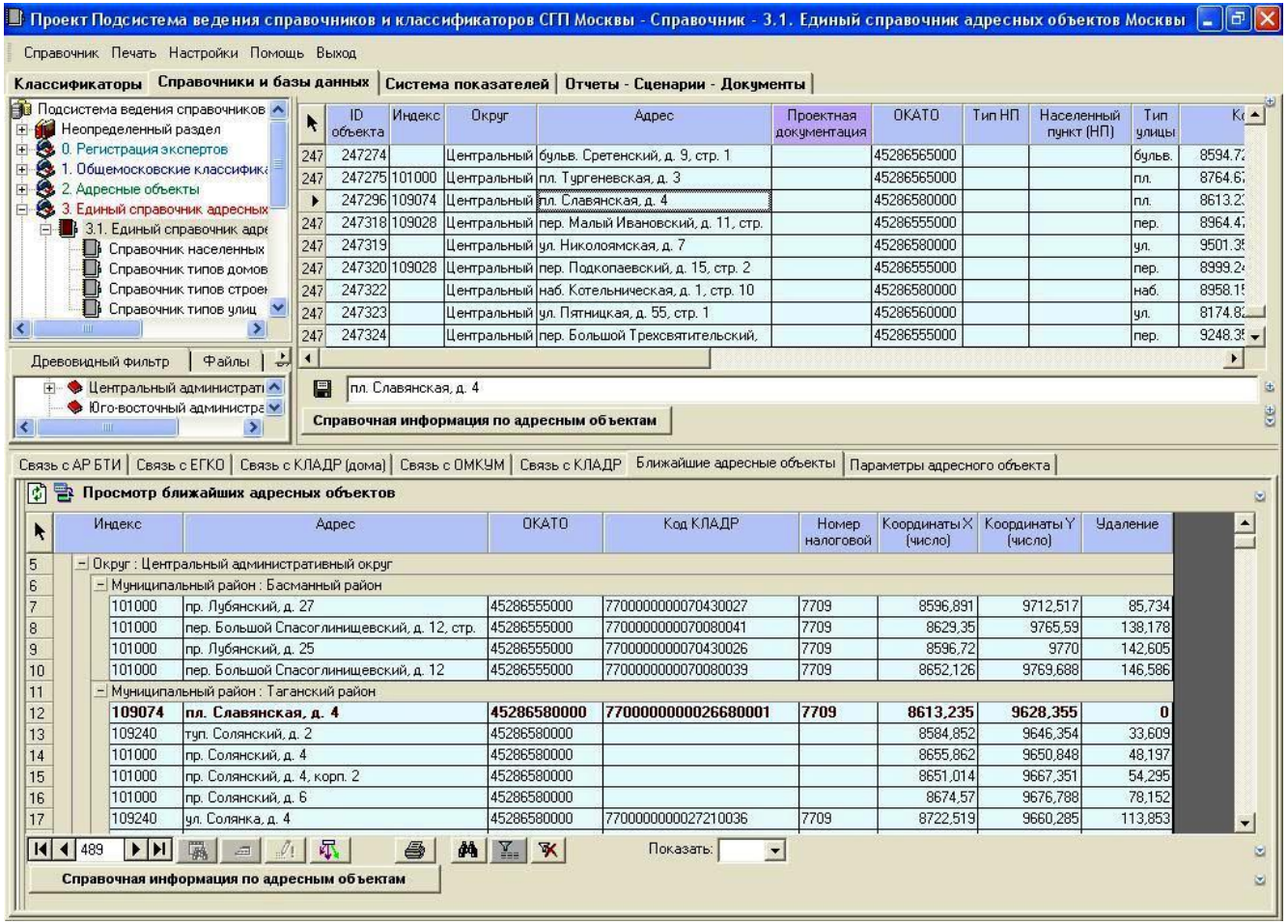
Результаты поиска

Поиск с использованием географических координат адресных объектов



Программное обеспечение поддерживает возможность поиска по координатам объекта или по его адресу (ArcGIS WGS84 - Bessel)

Поиск с использованием географических координат адресных объектов



Просмотр ближайших адресных объектов с координатным расчетом удаления

Нормализация адреса



	Код типа	Полное (ВР)	Полное (ВР-НР)	Полное (НР)	Аббревиатура (ВР)	Аббревиатура (НР)	Среднее с тчк (ВР)	Среднее с тчк (ВР-НР)	Среднее с тчк (НР)	Среднее бе тчк (ВР)
▶	01	АДМИНИСТРАТИВНЫЙ	Административный округ	административ	АО	ао	АДМ. ОКРУГ	Адм. округ	адм. округ	АДМ ОКРУГ
2	02	МУНИЦИПАЛЬНОЕ	Муниципальное	муниципальное	МО	мо	МУНИЦ.	Муниц. образ.	муниц. образ.	МУНИЦ
3	03	ТЕРРИТОРИАЛЬНАЯ	Территориальная единица	территориальна	ТЕ	те	ТЕР.	Тер. единица	тер. единица	ТЕР
4	04	ОБЛАСТЬ	Область	область			ОБЛ.	Обл.	обл.	ОБЛ
5	05	РАЙОН	Район	район			Р-Н.	Р-н.	р-н.	Р-Н
6	06	ГОРОД	Город	город			ГОР.	Гор.	гор.	ГОР
7	07	ДЕРЕВНЯ	Деревня	деревня			ДЕР.	Дер.	дер.	ДЕР
8	08	ПОСЕЛОК	Поселок	поселок			ПОС.	Пос.	пос.	ПОС
9	09	ПОСЕЛОК	Поселок городского типа	поселок	ПГТ	пгт	ПОС. ГОР.	Пос. гор. типа	пос. гор. типа	ПОС ГОР
10	10	ДАЧНЫЙ ПОСЕЛОК	Дачный поселок	дачный поселок			ДАЧН. ПОС.	Дачн. пос.	дачн. пос.	ДАЧН ПОС
11	11	АЭРОПОРТ	Аэропорт	аэропорт			АЭРП.	Аэрп.	аэрп.	АЭРП
12	12	СТАНЦИЯ	Станция	станция			СТАНЦ.	Станц.	станц.	СТАНЦ

Конфигуратор адресов позволяет настроить возможные видов представления в адресе основной и вспомогательной информации. Для этого настроены справочники компонентов адреса, необходимые для создания схем и вариантов написаний полного адреса объектов

Нормализация адреса

Перед запуском программы нормализации адресов необходимо вручную проконтролировать правильность «парсинга» - разборки адреса на составляющие компоненты (в случае работы с входным текстовым массивом), или – правильность заполнения отдельных элементов адреса (в случае работы с входным структурированным массивом).

Так, при текстовом написании адреса эксперту предлагается обработать следующие последовательные массивы:

Адреса, в которых не все данные разобраны

Адреса, для которых не указан индекс

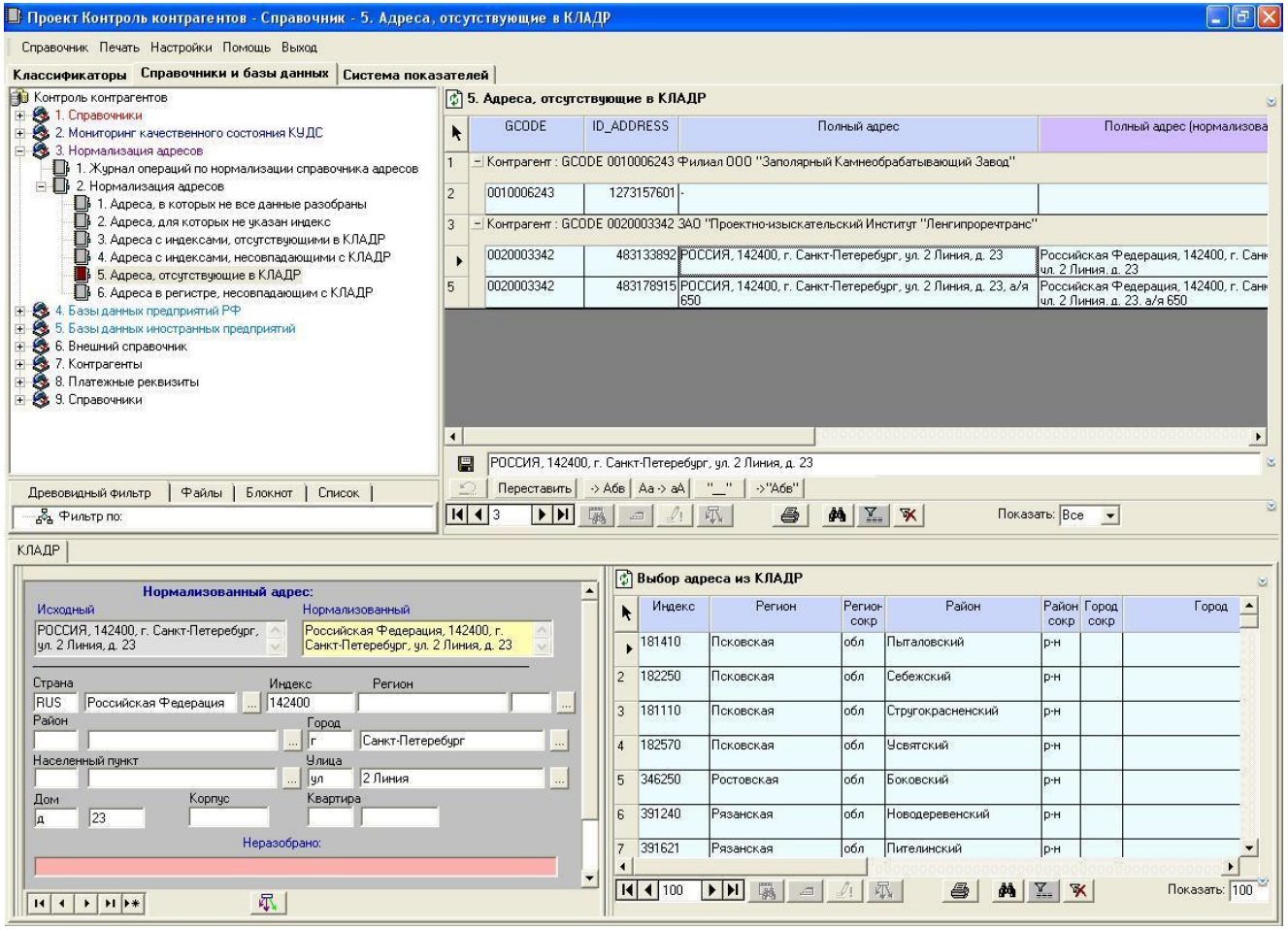
Адреса с индексами, отсутствующими в КЛАДР

Адреса с индексами, несовпадающими с КЛАДР

Адреса, отсутствующие в КЛАДР

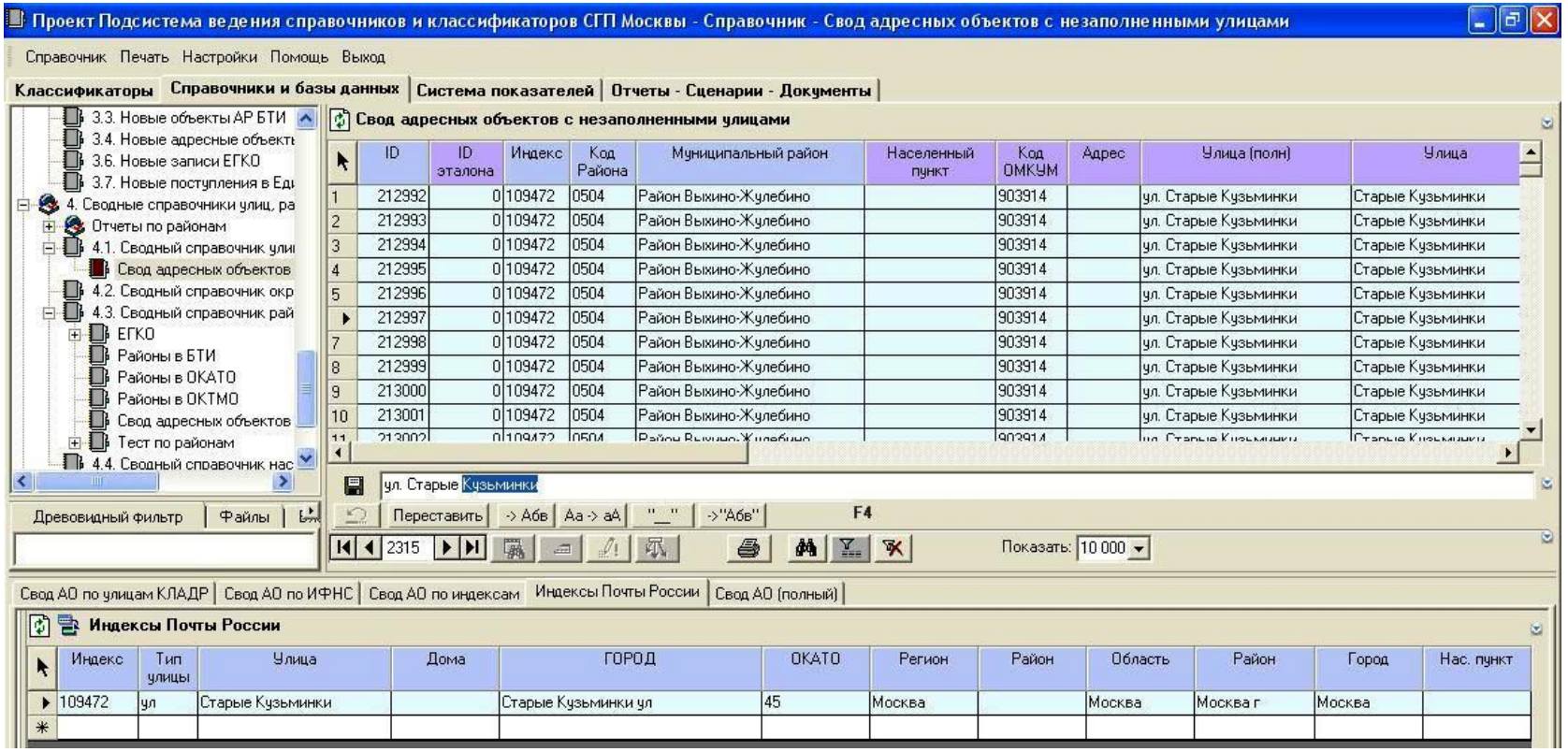
Адреса в регистре, несовпадающим с КЛАДР

Нормализация адреса



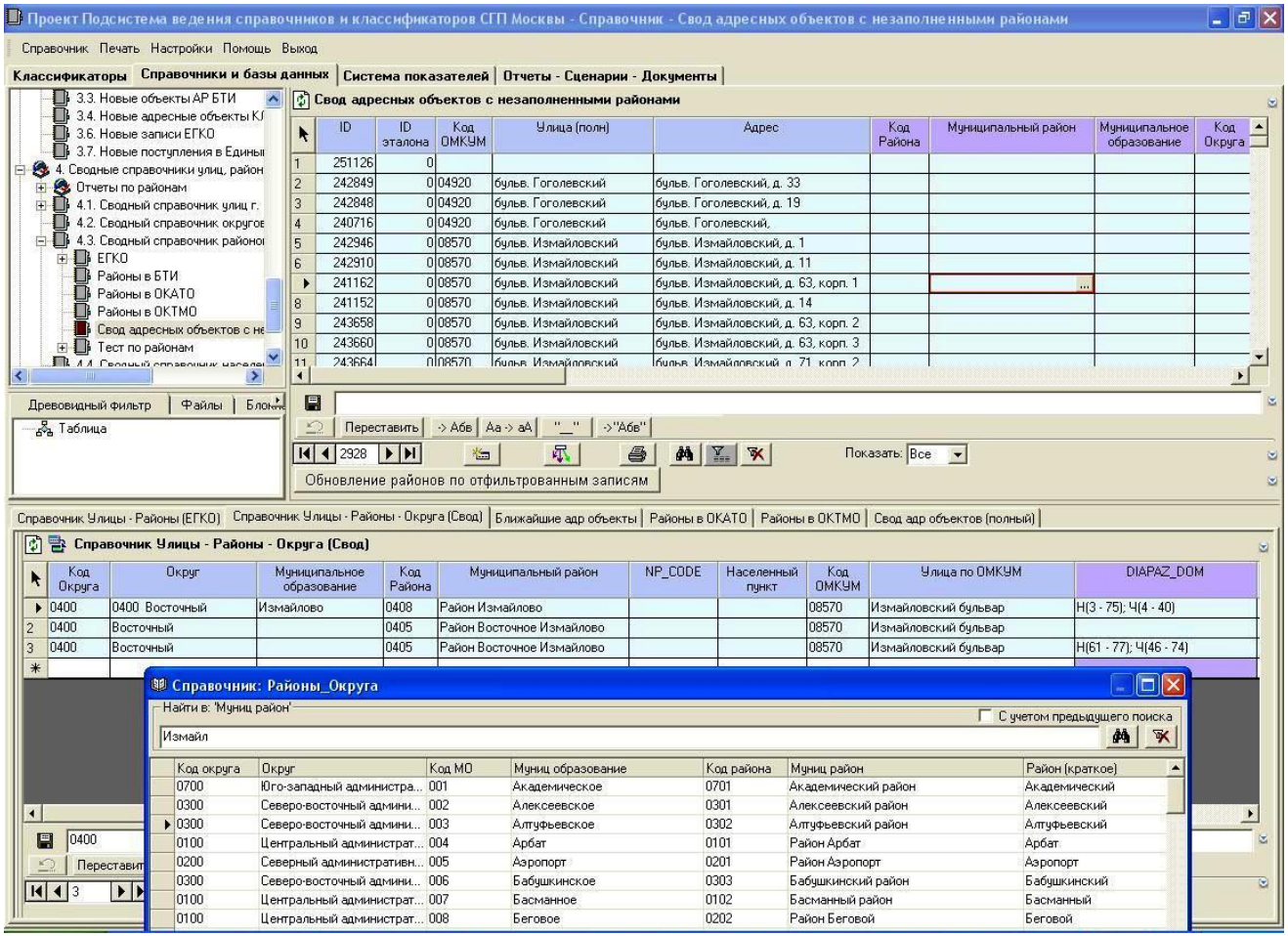
Выбор массивов

Нормализация адреса



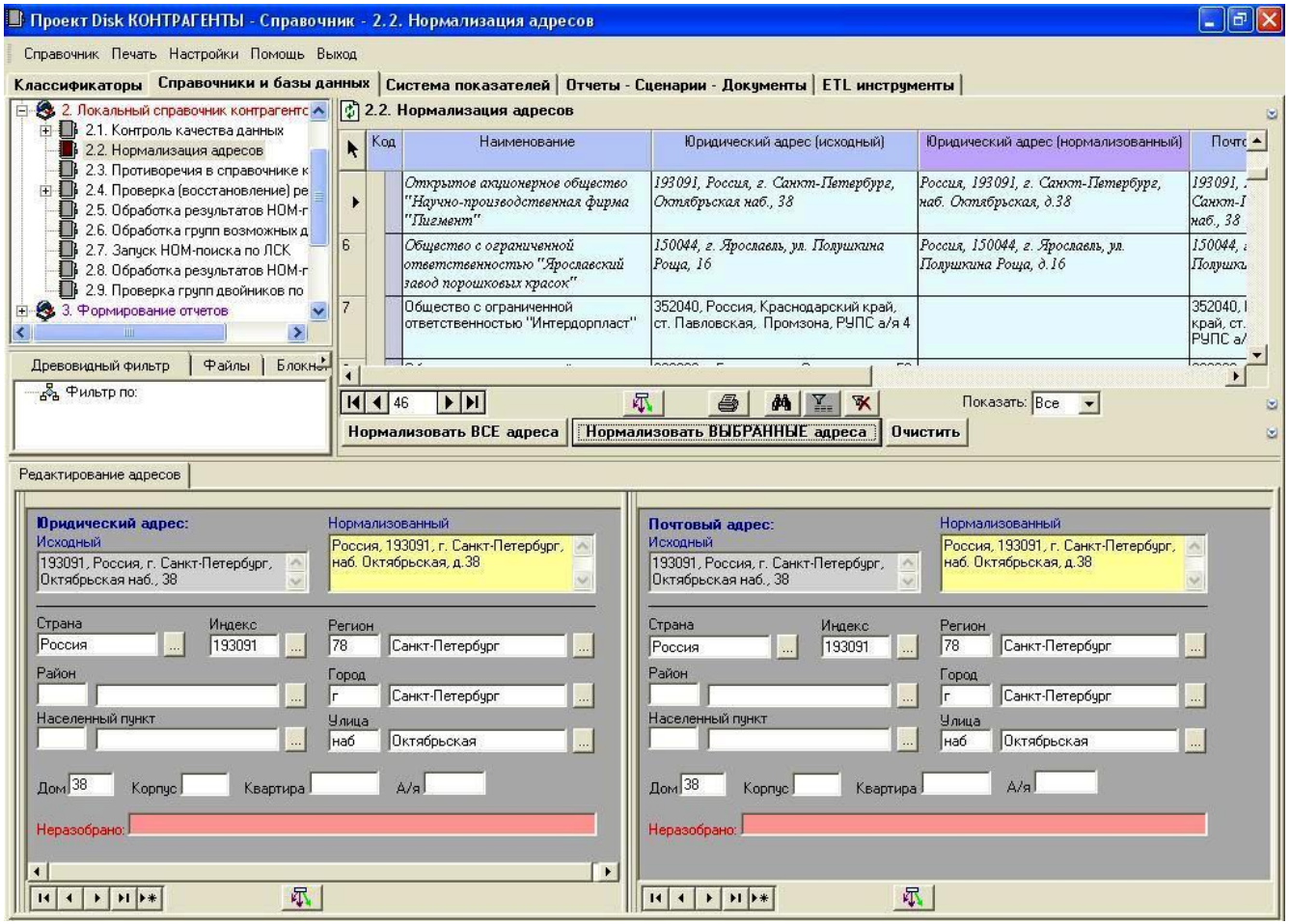
Пример формы проверки правильности названий улиц г. Москвы, отсутствующих в Общемосковском классификаторе

Нормализация адреса



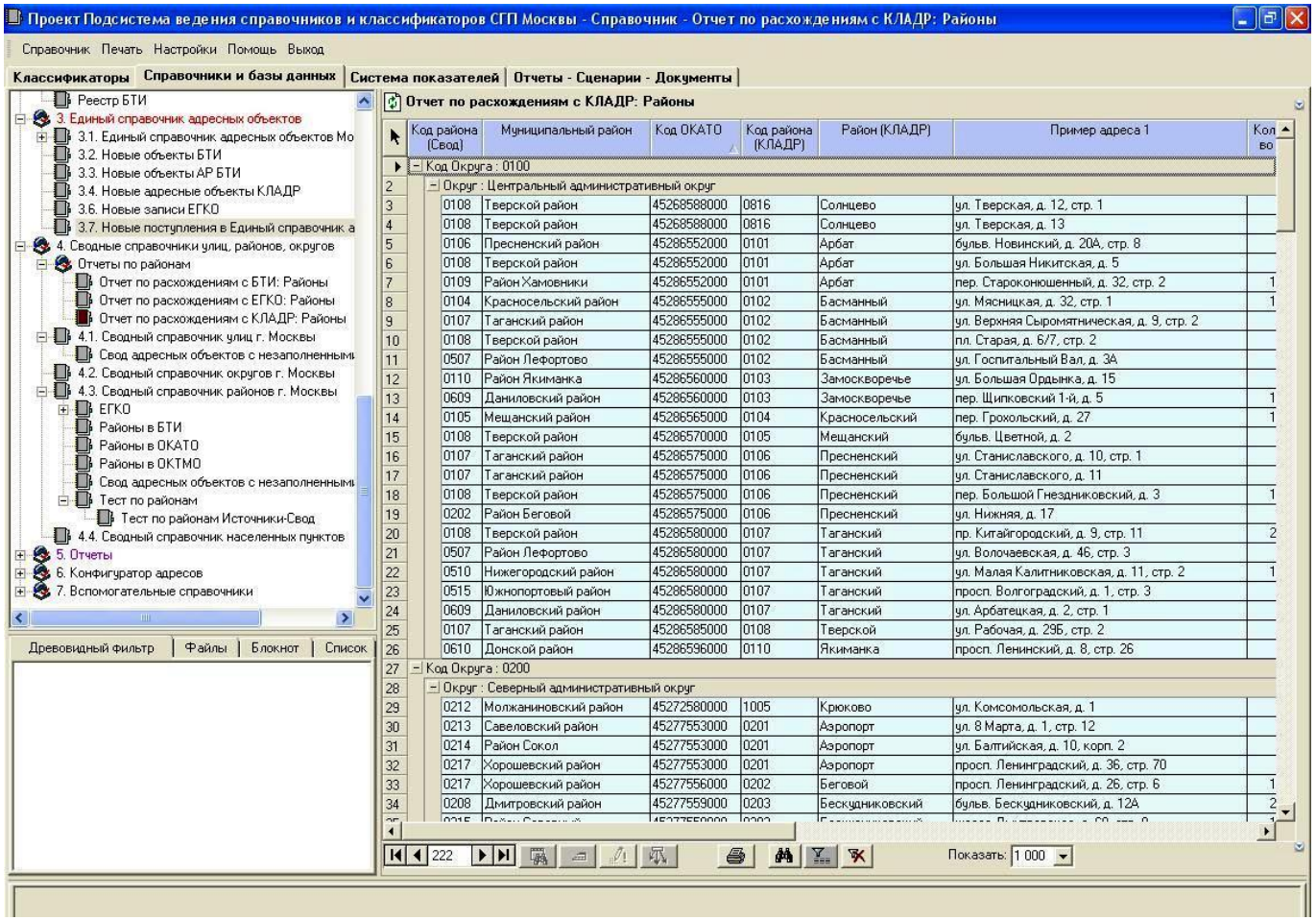
Пример формы проверки правильности присвоения адресным объектам кода района, в случаях, когда улица проходит по нескольким районам

Нормализация адреса



После выполнения всех экспертных проверок, запускается программа сборки нормализованного адреса

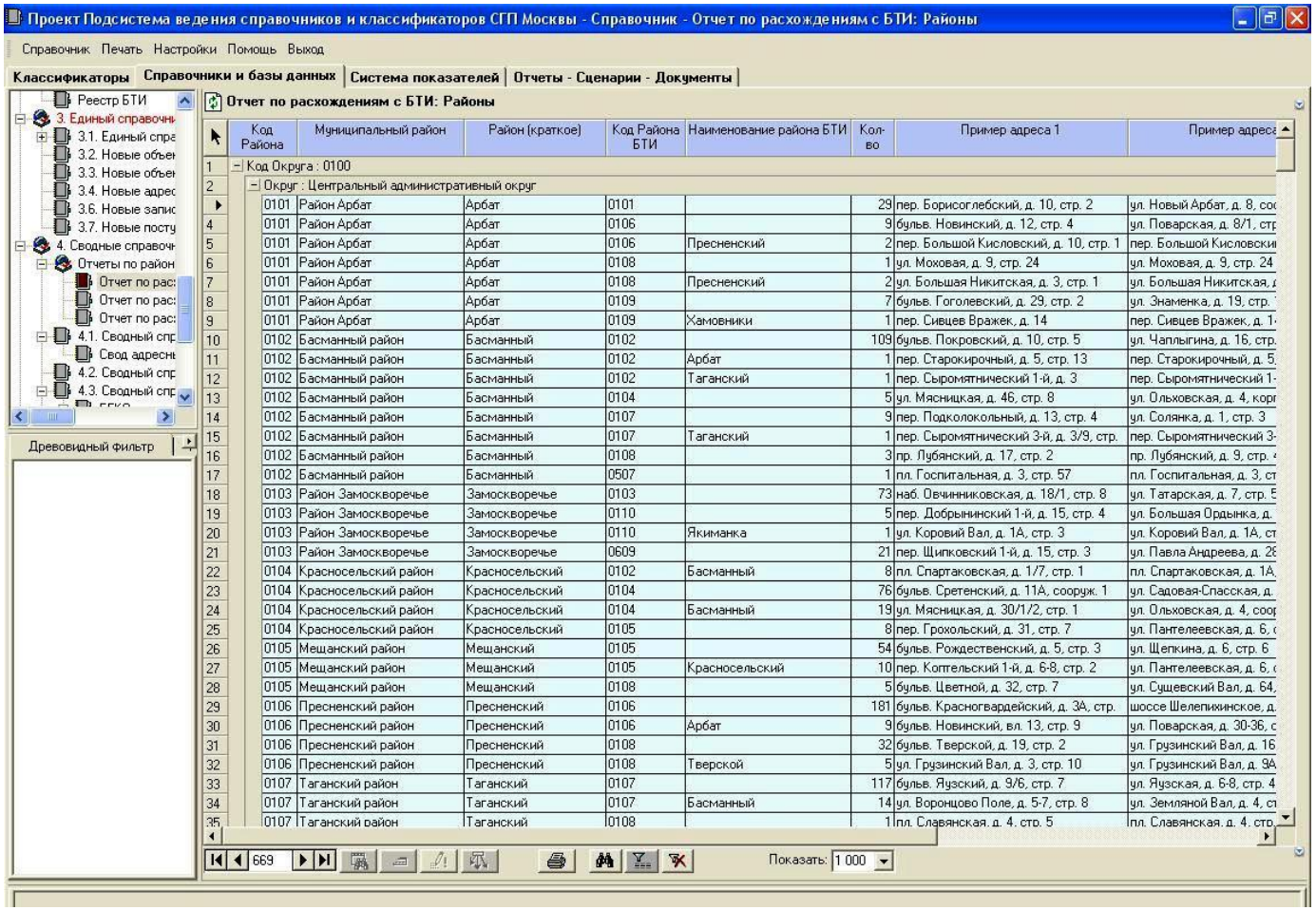
Выявление расхождений значений элементов адреса в различных источниках



	Код района (Свод)	Муниципальный район	Код ОКАТО	Код района (КЛАДР)	Район (КЛАДР)	Пример адреса 1	Кол во
	Код Округа : 0100						
2	Округ : Центральный административный округ						
3	0108	Тверской район	45268588000	0816	Солнцево	ул. Тверская, д. 12, стр. 1	
4	0108	Тверской район	45268588000	0816	Солнцево	ул. Тверская, д. 13	
5	0106	Пресненский район	45286552000	0101	Арбат	бульв. Новинский, д. 20А, стр. 8	
6	0108	Тверской район	45286552000	0101	Арбат	ул. Большая Никитская, д. 5	
7	0109	Район Хамовники	45286552000	0101	Арбат	пер. Староконюшенный, д. 32, стр. 2	1
8	0104	Красносельский район	45286555000	0102	Басманный	ул. Мясницкая, д. 32, стр. 1	1
9	0107	Таганский район	45286555000	0102	Басманный	ул. Верхняя Сыромятническая, д. 9, стр. 2	
10	0108	Тверской район	45286555000	0102	Басманный	пл. Старая, д. 6/7, стр. 2	
11	0507	Район Лефортово	45286555000	0102	Басманный	ул. Госпитальный Вал, д. 3А	
12	0110	Район Якиманка	45286560000	0103	Замоскворечье	ул. Большая Ордынка, д. 15	
13	0609	Даниловский район	45286560000	0103	Замоскворечье	пер. Щипковский 1-й, д. 5	1
14	0105	Мещанский район	45286565000	0104	Красносельский	пер. Грохольский, д. 27	1
15	0108	Тверской район	45286570000	0105	Мещанский	бульв. Цветной, д. 2	
16	0107	Таганский район	45286575000	0106	Пресненский	ул. Станиславского, д. 10, стр. 1	
17	0107	Таганский район	45286575000	0106	Пресненский	ул. Станиславского, д. 11	
18	0108	Тверской район	45286575000	0106	Пресненский	пер. Большой Гнездниковский, д. 3	1
19	0202	Район Беговой	45286575000	0106	Пресненский	ул. Нижняя, д. 17	
20	0108	Тверской район	45286580000	0107	Таганский	пр. Китайгородский, д. 9, стр. 11	2
21	0507	Район Лефортово	45286580000	0107	Таганский	ул. Волочаевская, д. 46, стр. 3	
22	0510	Нижегородский район	45286580000	0107	Таганский	ул. Малая Калитниковская, д. 11, стр. 2	1
23	0515	Южнопортовый район	45286580000	0107	Таганский	просп. Волгоградский, д. 1, стр. 3	
24	0609	Даниловский район	45286580000	0107	Таганский	ул. Арбатецкая, д. 2, стр. 1	
25	0107	Таганский район	45286585000	0108	Тверской	ул. Рабочая, д. 29Б, стр. 2	
26	0610	Донской район	45286596000	0110	Якиманка	просп. Ленинский, д. 8, стр. 26	
27	Код Округа : 0200						
28	Округ : Северный административный округ						
29	0212	Молжаниновский район	45272580000	1005	Крюково	ул. Комсомольская, д. 1	
30	0213	Савеловский район	45277553000	0201	Аэропорт	ул. 8 Марта, д. 1, стр. 12	
31	0214	Район Сокол	45277553000	0201	Аэропорт	ул. Балтийская, д. 10, корп. 2	
32	0217	Хорошевский район	45277553000	0201	Аэропорт	просп. Ленинградский, д. 36, стр. 70	
33	0217	Хорошевский район	45277556000	0202	Беговой	просп. Ленинградский, д. 26, стр. 6	1
34	0208	Дмитровский район	45277559000	0203	Бескудниковский	бульв. Бескудниковский, д. 12А	2

Пример выявленных расхождений с КЛАДР

Выявление расхождений значений элементов адреса в различных источниках



Пример расхождений с БТИ

Основные справочники
Вспомогательные справочники для систем поиска
Настройка систем ведения

Синхронизация справочников адресных объектов

Импорт и оценка качества данных в исходных справочниках

Создание и ведение Сводного справочника адресных объектов

Создание и ведение Сводных справочников вспомогательной информации

Организация работы экспертов

Аналитическая работа со сводными массивами НСИ, подготовка различных отчетов и выборок данных по заказам пользователей системы

Пронин Ю.Н.
Директор ЦНСИ
Компьютел
Pronin_1@mail.ru