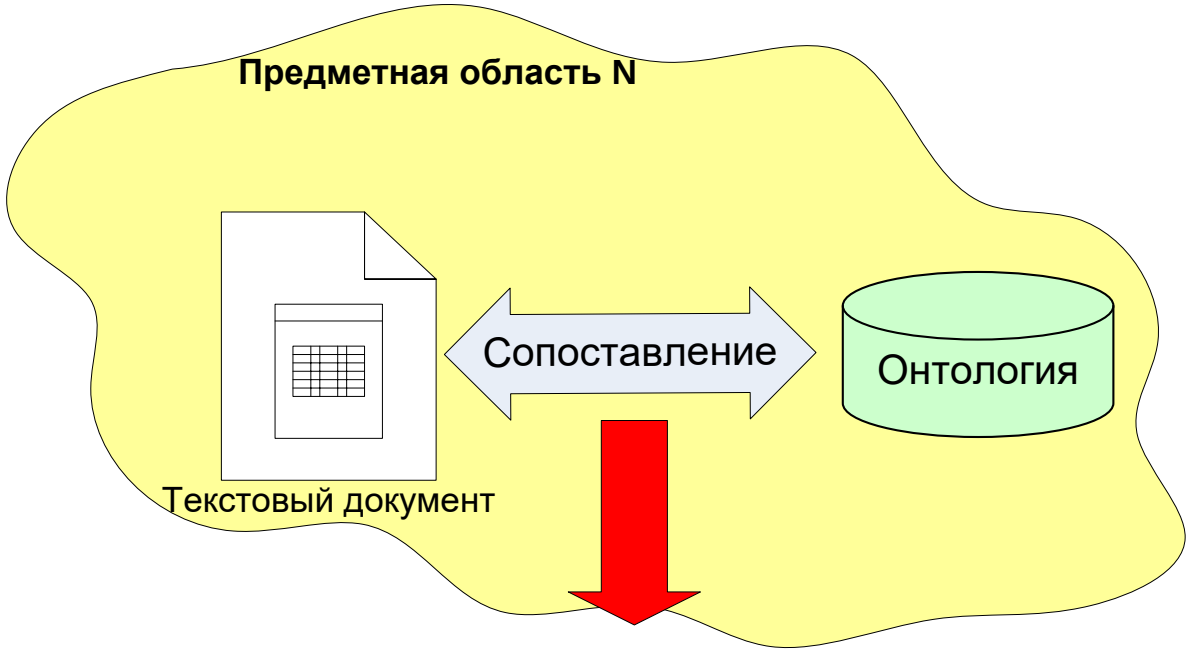
Занятие 5. Пример разработки информационной системы семантического контроля на основе требований ЕСПД.

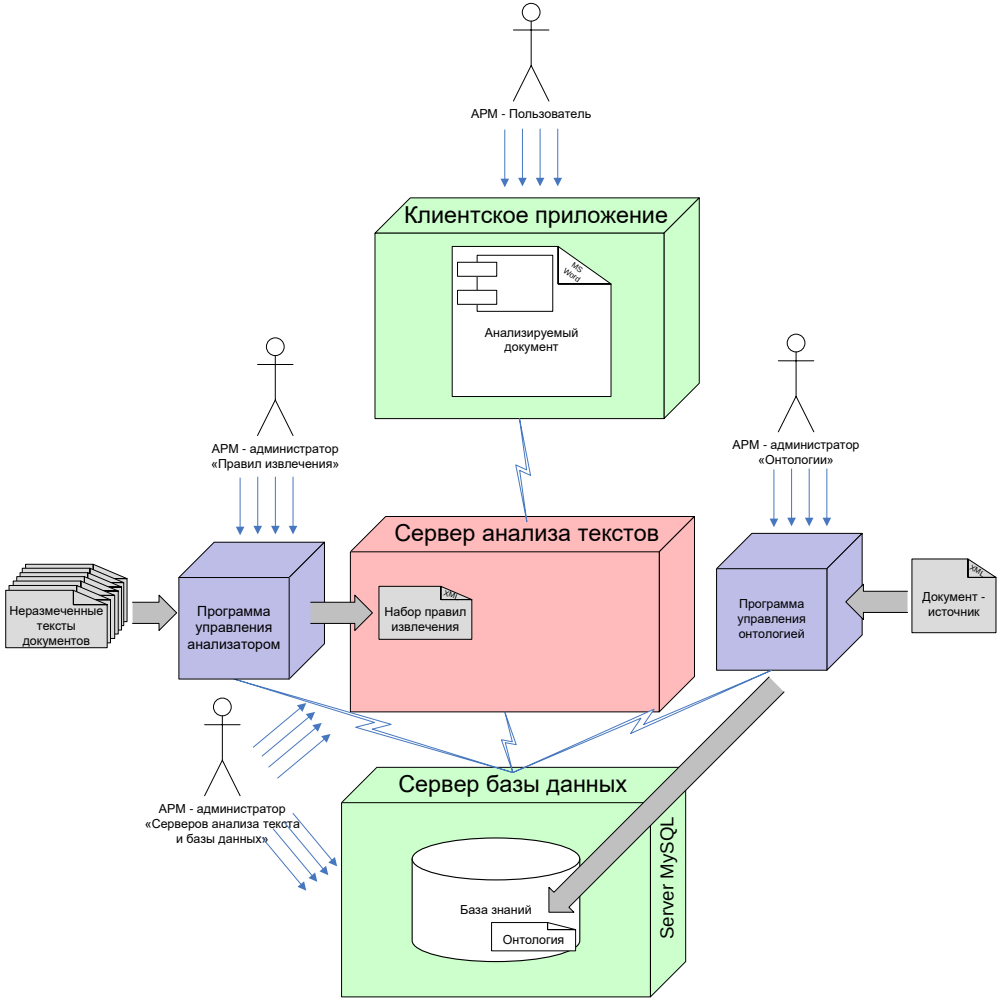
Интеллектуальная информационная система проверки и исправления почтовых адресов клиентов банка

ПОДХОД К РЕШЕНИЮ ЗАДАЧИ СЕМАНТИЧЕСКОГО КОНТРОЛЯ ТЕКСТОВ ДОКУМЕНТОВ



- Несоответствие 1,
- Несоответствие 2,
- Несоответствие 3,
...

ОБЩАЯ СТРУКТУРА СИСТЕМЫ СЕМАНТИЧЕСКОГО КОНТРОЛЯ



ИНТЕЛЛЕКТУАЛЬНАЯ ИНФОРМАЦИОННАЯ СИСТЕМА ПРОВЕРКИ И ИСПРАВЛЕНИЯ ПОЧТОВЫХ АДРЕСОВ КЛИЕНТОВ БАНКА

Система предназначена для автоматического выявления и исправления ошибок в почтовых адресах физических лиц на территории Российской Федерации и формирования «правильных» адресов в соответствии с классификатором адресов КЛАДР, который создан и ведется Федеральной налоговой службой России (http://www.gnivc.ru/inf_provision/classifiers_reference/kladr/).

Проверка почтовых адресов клиентов банка производится во взаимодействии с существующими информационными системами (ИС) банка. Почтовая информация ведется независимо несколькими ИС банка. Формат представления почтовых адресов определяется конкретной ИС банка и может варьироваться от жестко структурированного представления до представления в виде сплошной текстовой строки.

ОСНОВНЫЕ ФУНКЦИИ СИСТЕМЫ

Для ИС банка с жестко определенной структурой адреса разрабатываемая система выполняет следующие функции:

- выявляет и исправляет опечатки в наименованиях объектов, являющихся элементами адреса (город, улица и т.д.);
- проверяет реальное существование заданного адреса по классификатору КЛАДР;
- восстанавливает почтовый индекс адреса.

Для ИС банка со слабо определенной структурой адреса система дополнительно к перечисленным выше функциям выполняет выделение структуры адреса, т.е. составляющих элементов адреса (адресных объектов).

Система позволяет обеспечить:

- распознавание структуры исходного адреса, в том числе представленного сплошной строкой или группой сцепленных строк, и его представление в виде отдельных полей для адресных объектов;
- выявление и исправление опечаток и ошибок в адресных объектах;
- замену устаревших наименований адресных объектов, подвергшихся переименованию, на их актуальные наименования;
- восстановление почтового индекса в случае его отсутствия в исходном адресе или ошибочного написания;
- проверку существования почтового адреса по классификатору адресов КЛАДР;
- формирование выходного сообщения, содержащего структурированный исправленный адрес, а также информацию об ошибках;
- формирование протокола о результатах проверки почтового адреса.

ТИПОВЫЕ ОШИБКИ В ПОЧТОВЫХ АДРЕСАХ, ИСПРАВЛЯЕМЫЕ СИСТЕМОЙ

В процессе обработки входной информации **выявляются и исправляются следующие типовые ошибки в адресах**:

- орфографические ошибки в написаниях наименований регионов, населенных пунктов, улиц и т.д., вызванные ошибками правописания, ошибками при наборе с клавиатуры, шумовая и посторонняя информация;
- отсутствие или ошибочное заполнение почтового индекса;
- отсутствие наименований адресных объектов;
- отсутствие наименований типов адресных объектов;
- использование нестандартных, различных и неоднозначных сокращений для наименований типов адресных объектов: например, проезд - прд или пр., проспект - просп. или пр-т, ст. – станица или станция и др.;
- другие ошибки, возникающие при заполнении операторами форм ввода информацией о почтовом адресе клиента (например, часто возникают ошибки из-за неправильного выбора языка для ввода информации, ошибки в падежах, лишние знаки препинания и т.д.).
- выявляется наличие не существующих наименований адресных объектов.

РЕЖИМЫ РАБОТЫ СИСТЕМЫ

Разработанная система реализована в виде Web-сервиса, обеспечивающего единообразное функционирование системы в двух режимах: on-line и off-line.

- **Первый режим (on-line)** подразумевает прямое и оперативное взаимодействие между системой проверки почтовых адресов и ИС банка. Этот вариант отражает повседневную работу банка, когда в ИС заводятся учетные записи о новых клиентах или модифицируется информация о почтовых адресах у существующих клиентов. По данному сценарию ИС банка формируют запросы на проверку почтовых адресов, оформленных в виде XML сообщений. Система проверки адресов оперативно и в автоматическом режиме обрабатывает эти сообщения и в аналогичном формате возвращает результат проверки/исправления для последующей загрузки исправленного адреса в базу данных соответствующей ИС банка.
- **Второй режим (off-line)** используется, когда необходимо выполнить исправления группы адресов в пакетном режиме. В этом случае оператор выгружает порцию адресов конкретной ИС во временный файл и с АРМ оператора запускает процесс проверки адресов. В случае успешного завершения процесса проверки на экран АРМ оператора выдается протокол о результатах проверки, а также формируется ответный файл с набором проверенных и исправленных адресов. Получив результат обработки, оператор загружает исправленные адреса, обновляя содержимое БД конкретной ИС.

ИСПОЛЬЗУЕМЫЕ ТЕХНОЛОГИИ

Разработанная система **использует передовые технологии построения интеллектуальных систем**, оперирующих знаниями, ключевыми из которых являются технология извлечения знаний из неструктурированных данных и технология Semantic Web манипулирования знаниями.

Для распознавания адресных полей используются правила извлечения. Поскольку каждая ИС банка имеет собственное представление почтовых адресов, в разработанной системе заведены наборы правил извлечения, адаптированные под форматы каждой ИС. Правила извлечения хранятся в обычных XML файлах, так что в процессе эксплуатации существует возможность в случае изменения форматов представления почтовых адресов модифицировать правила извлечения без необходимости модификации самой системы.

Для проверки правильности адресов клиентов банка используется онтология географических наименований, построенная на основе классификатора КЛАДР. Онтология представляет собой семантическую сеть географических объектов, физически размещенную в реляционной базе данных.

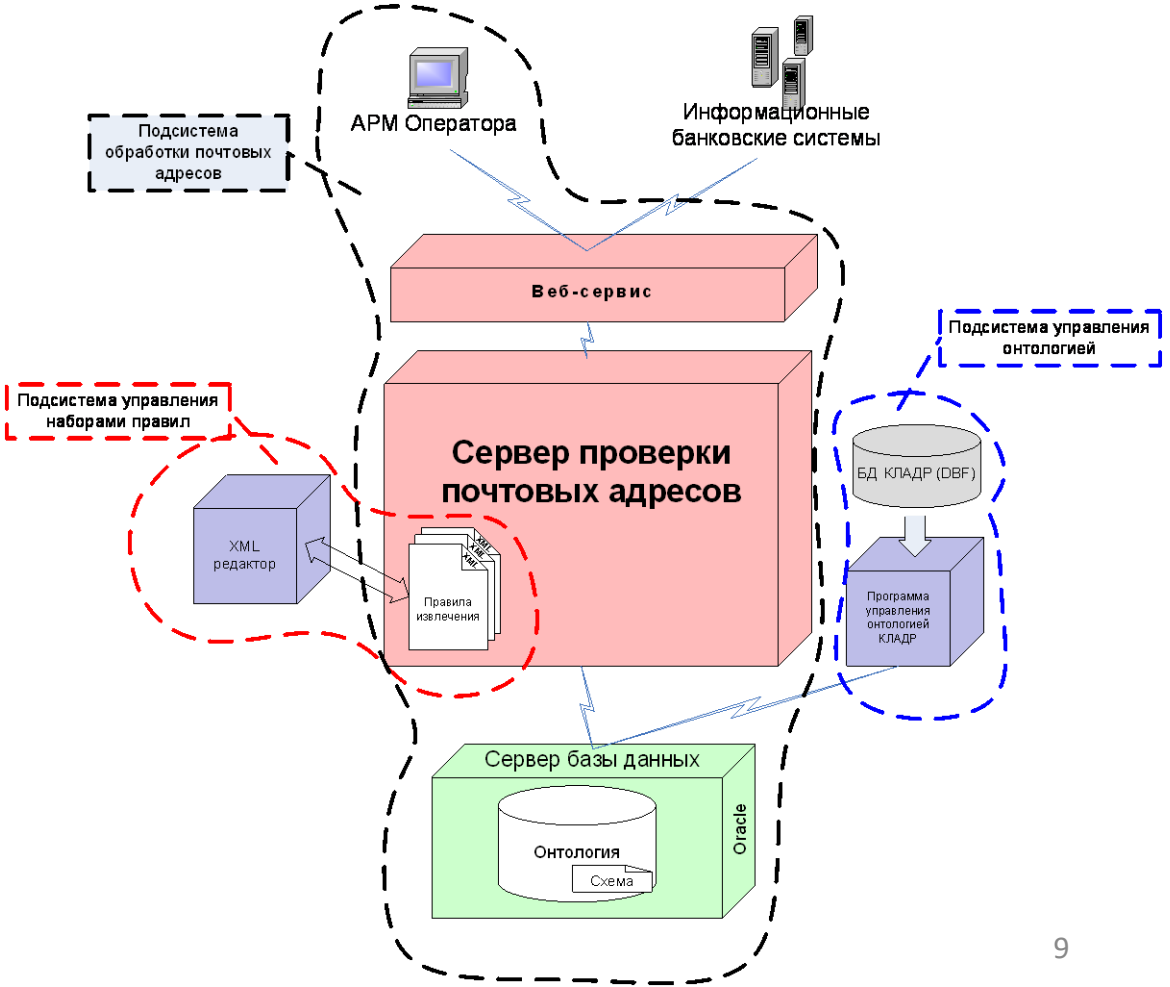
Для управления онтологией разработано специальное программное обеспечение, позволяющее, в частности, выполнять обновление данных на основе новых версий КЛАДР.

Кроме того, в системе используется разработанный авторами усовершенствованный метод обнаружения и исправления опечаток в названиях адресных объектов.

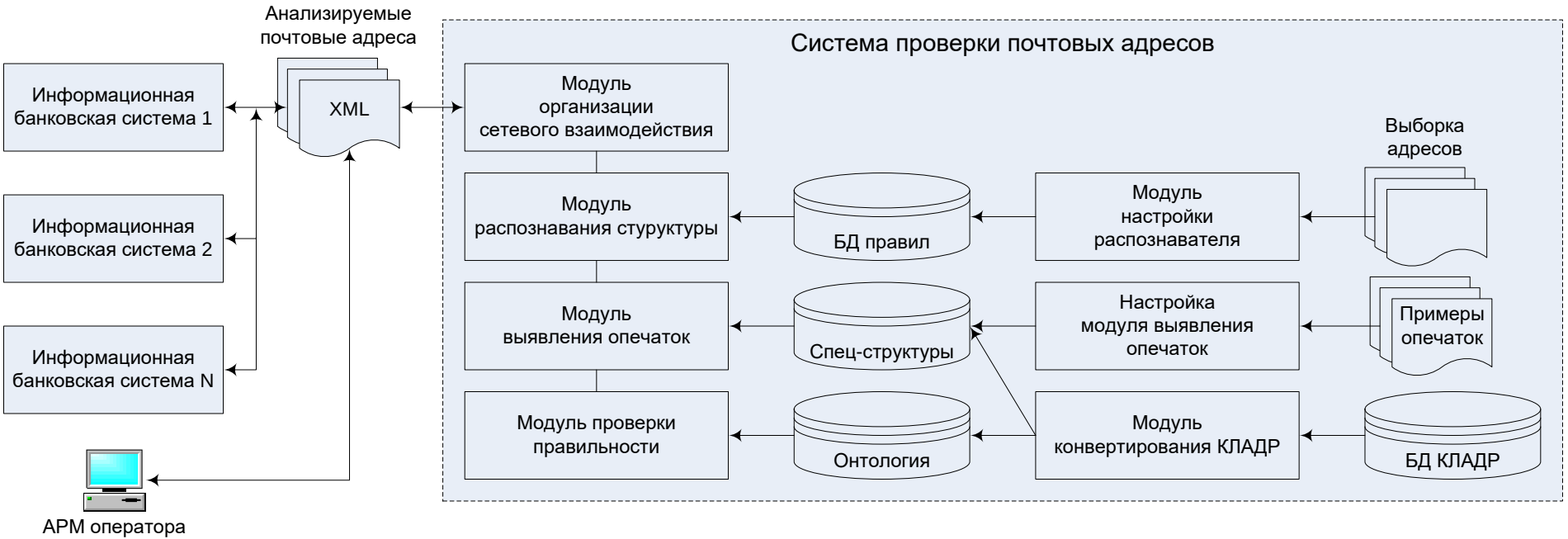
СТРУКТУРА СИСТЕМЫ

Система состоит из трех функциональных подсистем:

1. подсистема обработки почтовых адресов;
2. подсистема управления онтологией КЛАДР;
3. подсистема управления наборами правил.



СОСТАВ ПРОГРАММНЫХ МОДУЛЕЙ СИСТЕМЫ



ЭТАПЫ ЖЦ СИСТЕМЫ

При создании системы была выбрана **каноническая модель жизненного цикла**.
ЖЦ состоял из следующих стадий (этапов):

1. разработка и утверждение технического задания на создание ИС;
2. разработка программы и программной документации;
3. проведение предварительных испытаний;
4. проведение опытной эксплуатации;
5. проведение приемочных испытаний;
6. сопровождение ИС.

ЭТАП ОПЫТНОЙ ЭКСПЛУАТАЦИИ СИСТЕМЫ

После успешного завершения предварительных испытаний в течение года проводилась **опытная эксплуатация системы** на реальной информации и на программно-технических средствах банка.
Опытная эксплуатация ограничивалась взаимодействием с двумя ИС банка. При этом первая ИС банка являлась системой со слабо определенной структурой адреса, а вторая – с жестко определенной структурой адреса. В первом случае работа осуществлялась при функционировании системы в режиме off-line, во втором – в режиме on-line.

В процессе опытной эксплуатации решались следующие задачи:

- настройка системы на работу с конкретной ИС банка;
- проведение экспериментальных прогонов больших массивов реальных адресов в режиме off-line с целью проведения дополнительного тестирования и выявления более сложных ошибок в работе системы;
- проведение сравнительного анализа экспериментальной обработки больших массивов реальных адресов в режиме off-line с помощью разработанной системы и системы, используемой ранее для этих целей;
- проведение экспериментальных прогонов реальных адресов в режиме on-line с участием операторов, осуществляющих ввод информации о заёмщиках банка.

РЕЗУЛЬТАТЫ ОПЫТНОЙ ЭКСПЛУАТАЦИИ СИСТЕМЫ

В результате проведения экспериментальных прогонов больших массивов реальных адресов от первой ИС в режиме off-line выявлены более сложные ошибки в адресах (неправильные наименования типов населенных пунктов и типов улиц, неправильные составные наименования улиц, указание только многозначного номера дома вместо улицы, вместо наименований улиц, после слова «ул.» указано «отсутствует», «нет» или стоит знак «-» и др.), которые не исправлялись системой.

Сравнительный анализ экспериментальной обработки больших массивов реальных адресов (более 1000 адресов) в режиме пакетной обработки разработанной системой и ранее используемой для этой цели системой показал, что в целом разработанная система отбраковывает гораздо больше адресов (~18%), чем ранее используемая система (~4,5%), что отражает реализацию основной цели создания системы – не пропускать адреса, не соответствующие классификатору. Ранее используемая система пропускала много ошибочных адресов с точки зрения КЛАДР (~12%), в то время как разработанная система практически не пропускала ошибочные адреса (~0,001%).

Проведение экспериментальных прогонов реальных адресов в режиме on-line от второй ИС банка с участием операторов, осуществляющих ввод информации о заёмщиках показало значительно большее число отбракованных адресов (~25%) на массиве из 500 адресов по сравнению с аналогичным параметром для первой ИС. То есть оказалось, что система со слабо определенной структурой адреса подает на вход системы проверки и исправления адресов значительно более «чистую» информацию, чем система с жестко определенной структурой адреса.

АНАЛИЗ РЕЗУЛЬТАТОВ ОПЫТНОЙ ЭКСПЛУАТАЦИИ СИСТЕМЫ

Анализ проведенных экспериментов показал, что значительная часть ошибок на входе системы возникает из-за недостатков в информационном, программном и организационном обеспечении второй ИС банка:

- экранная форма, которую заполняют операторы, позволяет им в ряде случаев неоднозначно трактовать правила заполнение некоторых полей;
- для обработки адресной информации ИС банка используют внутренние классификаторы, которые не соответствует стандартизованным классификаторам и не актуализируются на постоянной основе службой эксплуатации;
- квалификация операторов оказывается в ряде случаев недостаточной для выполнения работы по оперативному заполнению входной формы с адресами клиентов банка.

Очевидно, что для повышения достоверности адресной информации о клиентах банка, кроме совершенствования самой системы проверки и исправления почтовых адресов, необходимо проведение комплекса работ по устранению вышеприведенных причин появления ошибок на входе этой системы.

Опытная эксплуатация системы показала её достаточно высокую эффективность по сравнению с ранее используемой системой пакетной проверки массива адресов. Система успешно исправляла от 75% до 90% всех поступающих на обработку адресов. Такой большой разброс обусловлен качеством поступающей на вход адресной информации, от правильности и полноты которой зависит качество работы системы в целом.