

**Техническое задание ОД-027 от 18 июня 2010 г.**  
**по приобретению программного комплекса для очистки, стандартизации и дедубликации контактной информации клиентов Банка по заявке № 6376**

1. Предметом тендера является наилучшее предложение по стоимости приобретения программного комплекса для очистки, стандартизации и дедубликации контактной информации клиентов Банка.
2. Подробные критерии к системе:
  - 2.1. Основные функции:
    - Профилирование данных - оценка качества данных по набору заранее заданных критериев и построение отчетов о качестве данных, позволяющих оценить состояние исходных данных.
    - Стандартизация данных - выполнение следующих операций:
      - очистка данных;
      - приведение данных к единому формату;
      - обогащение данных;
      - другие операции улучшения качества данных.
    - Дедубликация данных - применение специальных методов для выявления записей-дубликатов.
    - Консолидация данных - создание мастер-записи на основе выявленных записей-дубликатов.
    - Мониторинг - получение отчетов о текущем качестве данных и об изменении качества данных в течение промежутков времени.
  - 2.2. Достоверность изменений:
    - Система должна классифицировать данные по достоверности выполненных изменений и, в частности, позволять выделить множество гарантированно правильных изменений.
    - Гарантированно правильные изменения — это изменения контактных данных, в которых допускается не более 1 (одной) ошибки на 10 000 записей (99,99% изменений корректны).
    - При этом выявленные ошибки в данных, которые помечены как гарантированные, считаются ошибками системы 1-го приоритета и исправляются подрядчиком (производителем) в течение 24 рабочих часов.
  - 2.3. Система должна обеспечивать проведение следующих операций над клиентскими данными:
    - ФИО:
      - Разбиение ФИО по компонентам (фамилия, имя, отчество).
      - Определение пола по фамилии, имени и отчеству.
      - Выявление неполных ФИО с инициалами, извлечение инициалов.
      - Исправление опечаток в именах, отчествах.
      - Замена уменьшительно-ласкательных имен полными.
      - Обработка ФИО, написанных транслитом (при условии покупки модуля «Обратная транслитерация»).
      - Выявление несуществующих имен (например, «Не знаю»).
      - Простановка кода проверки распознанного ФИО.
    - Паспортные данные:
      - Проверка номера документа (паспорта, водительского удостоверения) по нормативам.

- Проверка серии документа (паспорта, водительского удостоверения) по нормативам.
  - Проверка учреждения, выдавшего паспорт.
  - Проверка кода подразделения паспорта.
  - Проверка даты выдачи паспорта.
  - Проверка срока действия паспорта по дате выдачи и дате рождения.
- Адресная информация:
    - Разбиение адреса по компонентам, включая: исправление опечаток и сокращений в адресе, приведение компонентов адреса к единому формату, проверку существования адреса, восстановление пропущенных компонентов адреса.
    - Обработка адресов, написанных транслитерацией.
    - Замена старых названий объектов на новые, например, «Ленинград» → «Санкт-Петербург».
    - Использование набора городов «по умолчанию» в случае, если в адресах пропущен город, но при этом известно местоположение клиентов (например, Москва и Мурманск).
    - Использование КЛАДР и при желании дополнительных по отношению к КЛАДР справочников для распознавания адресов не из КЛАДР (Мосреестр, распространенные улицы и пр).
    - Использование справочника Мосреестр для распознавания московских строений и определения типов строений (жилое/нежилое).
    - Определение кода КЛАДР.
    - Простановка кода качества для распознанного адреса.
    - Простановка кода проверки распознанного адреса.
    - Протоколирование причин изменений адреса.
    - Подстановка индексов, регионов, населенных пунктов по номеру телефона или по КЛАДР.
    - Проверка адресной информации на корректность.
  - Информация о телефонах:
    - Выделение кода города из телефона.
    - Приведение компонентов телефона к единому формату.
    - Проверка существования телефонного кода.
    - Проверка и восстановление телефонных кодов на основании почтового адреса.
    - Преобразование региональных телефонных кодов в федеральные.
    - Разделение домашних, рабочих, мобильных номеров по разделителям, например «домашний», «сотовый», «рабочий».
    - Выделение мобильных телефонов.
    - Выявление несуществующих телефонов (например, «11111111»).
    - Проверка соответствия длины телефона системе нумерации населенного пункта.
    - Разделение множества телефонов, заданных одной строкой, например, «моб 916 1510679, дом 320-78-10, рабочий 80951128912 доб. 3342» с учетом их типа.
    - Замены телефонных кодов для устаревших телефонов.
    - Простановка кода проверки распознанного телефона.
  - Адрес электронной почты:

- Проверка синтаксиса e-mail.
  - Проверка состояния домена, к которому принадлежит e-mail.
  - Корректировка с учетом распространенных ошибок, которые допускают пользователи при вводе e-mail.
- 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.
- 2.4. Дедубликация:
- Система должна обеспечивать использование различных методов поиска дубликатов.
    - Строгие методы - методы сравнения, при которых дубликаты ищутся на основе точного совпадения соответствующих полей.
    - Вероятностные методы - методы сравнения, когда решение о наличии дубликата принимается на основе вероятности схожести полей у двух записей. В системе должны быть реализованы не менее 5 различных алгоритмов вероятностного сравнения с простым выбором между ними.
  - Система должна обеспечивать автоматическое выделение мастер-записей из наборов дубликатов записи и её обогащение данными из других записей.
3. В комплект поставки решения должны входить словари и справочники, обеспечивающие выполнение вышеуказанных операций. Обязательным условием поставки является предоставление услуги по регулярному обновлению словарей и справочников. Система должна позволять подключать собственные словари и справочники Банка, а также редактировать словари и справочники.
4. Система должна обеспечивать не менее 85% качественных записей по каждому полю после проведения обработки при условии, что записи изначально содержат минимальную, необходимую для восстановления информацию.
5. Система должна быть универсальной и работать с любыми типами данных, такими как данные о клиентах, продуктах, банковские реквизиты и т.д.
6. Пользовательский интерфейс решения должен удовлетворять следующим требованиям:
- Система должна обладать интуитивно понятным интерфейсом среды разработки, в работе с которым не требуются специальные навыки программирования для разработки процессов силами специалистов Банка.
  - Разработка процессов обеспечения качества данных должна осуществляться без программирования путём реализации логики преобразований.
  - Система должна обеспечивать быстрое внесение изменений в уже разработанные процессы обеспечения качества данных.
  - Система должна позволять разрабатывать новые процессы и сопровождать существующие силами специалистов заказчика.
  - Система должна генерировать следующие отчеты:
    - Отчет о текущем качестве данных по полям
    - Отчет об историческом изменении качества данных
7. Требования к обработке данных:

- Система должна позволять использовать внешние функции и процедуры.
  - Система должна позволять запускать процессы обработки данных следующими способами:
    - По расписанию;
    - По запросу;
    - С использованием командной строки;
  - Система должна обеспечивать проведение ручной обработки данных оператором для следующих ситуаций:
    - Невозможно автоматически обработать запись;
    - Невозможно автоматически определить являются ли записи дубликатами или нет;
    - Невозможно автоматически создать обогащённую мастер-запись.
8. Система должна обеспечивать подключение к базам данным СУБД (ORACLE, Teradata), и подключение к различным источникам через ODBC, а также работу с плоскими и XML файлами.
9. Система должна удовлетворять следующим требованиям к производительности:
- Выполнять стандартизацию, дедупликацию, консолидацию не менее 100 000 клиентских записей в час на Стандартной конфигурации оборудования.
  - Система должна обеспечивать линейную зависимость производительности и времени обработки информации от аппаратного обеспечения и количества записей.
10. Система должна обеспечивать журналирование процессов обеспечения качества и интеграции данных.

*30 июля 2010 г.*

**Директор ОД ИТД**

**Виноградов Е.А.**