Консолидация информации о субъектах

© Татьяна Лякишева

ООО «ФОРС-Центр Разработки»
Tlyakish@fors.ru

Аннотация

Данная работа посвящена вопросам построения реестров юридических и физических лиц на основе данных из различных источников.
Описывается технология формирования таких реестров путем консолидации сведений из баз данных – источников, а также принципы сопоставления, идентификации и единообразной интерпретации данных.
Рассматриваются подходы, примененные в рамках реальных промышленных проектов.

1. Задачи и проблемы консолидации сведений о субъектах из различных источников

Большинство проектов по созданию корпоративных и ведомственных информационных систем требуют создания интегрированных электронных реестров субъектов (физических и юридических лиц). Информация о субъектах, как правило, фрагментирована и рассредоточена по различным источникам, и требуется ее корректная консолидация для обобщенного представления потребителям.
Предпосылками для консолидации сведений о субъектах из различных источников являются:

- Наличие в различных информационных ресурсах (ИР) данных о субъектах, не совместимых семантически – дублирование информации в различных ИР, риск расхождений при консолидации информации.
- Отсутствие эталонной модели данных о субъектах, отсутствие полной картины имеющейся информации.
- Необходимость создания информационного ядра для интеграции ИР.

Рассмотрим данную проблему на примере федеральных исполнительных органов. Проблема интеграции информационных ресурсов на сегодня актуальна для большинства из них. В качестве источников в первую очередь рассматриваются базы данных, содержащие информацию о субъектах - физических и юридических лицах. Это общероссийские информационные ресурсы, такие как Единый Государственный Реестр Юридических Лиц (МНС), Государственный реестр предприятий и организаций (Госкомстат), а также ведомственные базы данных.

Целью интеграции является предоставление потребителям (пользователям) максимально полных и всесторонних сведений о множестве субъектов, находящихся в определенных правоотношениях с соответствующим исполнительным органом (например, субъектах регулирования). Для этого в той или иной форме должна решаться задача идентификации, однако ее решение осложняется следующими обстоятельствами:

- Отсутствует единый поисковый и идентифицирующий атрибут (код) для физических и для юридических лиц. В случае российских юридических лиц, на его роль отчасти претендует ОГРН, однако во многих случаях он не может считаться уникальным идентификатором субъекта.
- Для идентификации юридических лиц в разных базах данных могут использоваться ИНН (+КПП), ОКПО, ОГРН. Однако отсутствует единая эталонная федеральная база данных, где были бы корректно заполнены и сопоставлены друг другу все эти коды. Более того, для юридических лиц степень детализации в различных учетных системах-источниках различна, и при взаимном сопоставлении могут устанавливаться связи один-ко-многим.
- Для идентификации российских физических лиц в отсутствие идентификатора персональных данных могут быть использованы атрибуты документов, однако в связи с большим количеством типов документов, удостоверяющих личность, и проведенной заменой паспортов эти атрибуты не достаточны для проведения качественной идентификации.
- Для физических и юридических лиц – нерезидентов в ведомственных базах данных в принципе отсутствуют достоверно и однозначно идентифицирующие их реквизиты.

Труды 8ой Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL'2006, Суздаль, Россия, 2006.

• Отсутствуют либо не соблюдаются единые стандарты на заполнение текстовых полей (место рождения ФЛ, адреса, информация о выдаче документа удостоверяющего личность), что затрудняет их использование для поиска и идентификации.
• Общая проблема - низкое качество заполнения баз данных, ведение которых территориально распределено и производится с помощью различных программных средств.

При решении задач интеграции сведений могут быть использованы принципиально различные подходы: от интеграции информации в режиме реального времени (EII – Enterprise Information Integration) до пакетной интеграции данных (ETL – Extract, Transform, Load).

Технология EII лучше всего подходит в тех случаях, когда необходимо создать общий шлюз (gateway) с едиными языком и точкой доступа к несогласованным источникам данных. Она применима тогда, когда целью интеграции является выдача всесторонней информации о субъекте по запросу. При этом пользователь вынужден максимально расширять круг поиска и количество проверяемых ресурсов, чтобы не пропустить существенную информацию. Как следствие, процесс получения полной и всесторонней информации о субъекте из различных информационных ресурсов каждый раз требует ручного «просеивания» полученных данных и становится достаточно трудоемким.

Более качественных результатов в общем случае можно достичь за счет предварительного формирования эталонных реестров лиц путем автоматической (или автоматизированной) обработки сведений из баз данных – источников. Этот путь (ETL) позволяет на этапе загрузки реестра провести очистку и идентификацию данных, и сформировать эталонное представление сведений о субъектах – так называемые «главные наборы данных». Именно этот подход применяется в ряде проектов, реализованных ООО «ФОРС-Центр Разработки» для федеральных ведомств. Разработанные системы строятся на общем программном ядре, с использованием единых принципов проектирования и кодирования, на основе повторного использования прикладных и технологических компонент системы. Это позволяет говорить о сформировавшейся программной платформе, получившей название ВЕРО.

2. Что такое ВЕРО

Программная платформа ВЕРО («Ведение единого реестра объектов») представляет собой готовое программное ядро системы, реализующее функции, общие для большинства информационных систем масштаба организации, а также совокупность шаблонов, компонент, методов построения на основе этого ядра полнофункционального программного решения. Архитектура ВЕРО основана на концепции ведения единого реестра объектов прикладной области.
В технологическом ядре информационной системы, построенной на основе ВЕРО, можно условно выделить следующие «слои»: единый реестр объектов (объектное ядро), основные технологические компоненты, прикладные компоненты (см. рис. 1).

Рис. 1. Технологическое ядро системы.

В частности, одним из типовых решений в составе ВЕРО является программная реализация процесса формирования консолидированного реестра персон (юридических и физических лиц) на основе сведений из баз данных – источников информации. Эта компонента получила название Картотека. В следующих разделах приведено общее описание применяемых в Картотеке принципов и подходов.

3. Организация реестра субъектов

Принципы интеграции, реализованные в Картотеке, основаны на выделении фиксированной группы идентифицирующих и поисковых атрибутов для физических и юридических лиц. Эти атрибуты сохраняются в записях-«упоминаниях», после чего производится идентификация их с другими упоминаниями в различных источниках.
Под упоминанием подразумевается информационный объект определенной структуры, содержащий основные атрибуты, характеризующие лицо (их зачастую называют установочными реквизитами), преобразованные к

единому представлению. Кроме того, упоминание содержит информацию о своем источнике (информационном ресурсе, из которого оно поступило) и уникальный идентификатор в источнике, позволяющий однозначно извлечь исходную запись из БД источника. Каждое упоминание само по себе может являться основанием для регистрации в реестре нового субъекта – соответственно юридического или физического лица. Так как упоминания единообразно представляют данные о субъектах, они могут быть сопоставлены между собой (идентифицированы). При выявлении соответствия, упоминания объединяются: устанавливается их принадлежность одной и той же карточке субъекта. Карточка субъекта (или просто субъект) является информационным объектом, содержащим набор атрибутов (установочных реквизитов лица), сформированных по упоминаниям (более подробно об этом будет рассказано далее). Карточки субъектов и упоминания связаны реляционным отношением 1**:** N. Множество карточек субъектов составляет Картотеку.

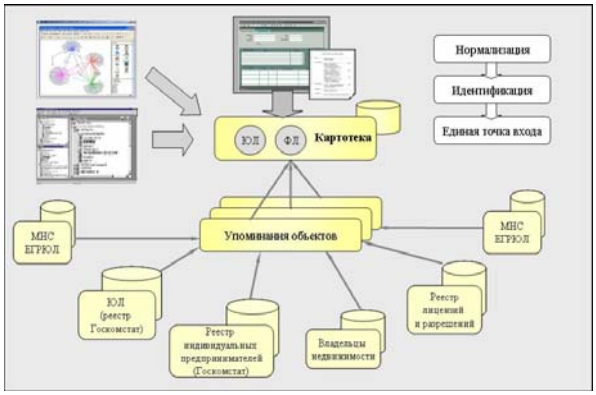

Рис. 2. Схема организации Картотеки.

В качестве установочных реквизитов могут быть использованы:

- для физических лиц - фамилия, имя, отчество, ИНН, дата и место рождения, реквизиты документов, адреса регистрации и места жительства;
- для юридических лиц - полное и краткое наименование, ИНН, КПП, ОКПО, ОГРН, юридический и фактический адреса.

По сути, Картотека представляет собой ядро системы электронных досье, в которой может быть обеспечен «сквозной» поиск по значениям основных реквизитов, а для выбранного лица – прямой доступ к информации о нем в базах данных – источниках информации.

В процессе формирования Картотеки последовательно решаются следующие задачи:

- приведение исходных данных о субъекте, содержащихся в системе-источнике, к унифицированному представлению для получения упоминаний (нормализация);
- сопоставление упоминаний между собой и установление их соответствия карточкам субъектов (идентификация).

4. Нормализация атрибутов

Под нормализацией в контексте Картотеки подразумевается преобразование отдельных атрибутов к некоторому каноническому виду с учетом их семантической нагрузки. В различных базах данных могут быть заданы различные форматы представления одних и тех же реквизитов. В качестве примеров реквизитов, как правило, требующих нормализации, можно привести:

- Фамилию, имя и отчество физического лица. В некоторых базах данных эта информация помещается в одно поле, причем встречается разный порядок и степень детализации при заполнении (фамилия – имя - отчество, имя – отчество - фамилия, фамилия - инициалы, инициалы – фамилия). При подключении к Картотеке необходимо реализовать отображение существующей структуры данных (или одного строкового поля) в структуру упоминания, содержащую в частности, атрибуты: фамилия, имя, второе имя, отчество, инициалы (первые буквы имени и отчества).
- наименование организации. В различных случаях в БД источников могут использоваться: краткое наименование, краткое наименование с указанием организационно-правовой формы, полное наименование, наименование на иностранном языке. Поле или совокупность полей, содержащих эту информацию, должна быть преобразована в структуру упоминания, содержащую в частности, атрибуты: код организационно-правовой формы (по общероссийскому классификатору ОКОПФ), сокращенное наименование, полное наименование.
- адрес – пожалуй, самый очевидный случай, когда реквизит требует нормализации. Как правило, адрес вносится в одно текстовое поле. При ручном заполнении БД, в отсутствие жесткого стандарта написания адресов, значения строк становятся несопоставимы, и не могут быть использованы для идентификации и поиска субъектов. Пример (заполнение поля адреса в одной и той же базе данных) приведен в таблице.

Таблица. Пример заполнения поля адреса

693000 Г Ю САХАЛИНСК ПРОСПЕКТ МИРА 1А
693000 Г ЮЖНО САХАЛИНСК ПР МИРА 1 А
693000 Г.ЮЖНО-САХАЛИНСК ПРОСП.МИРА 1 А
Г.Ю-САХАЛИНСК, ПР.МИРА 1А
САХАЛИНСКАЯ ОБЛ. Г.Ю-САХАЛИНСК, ПР.МИРА 1 А
САХАЛИНСКАЯ ОБЛ. Г.Ю-САХАЛИНСК, ПР.МИРА 1А
САХАЛИНСКАЯ ОБЛ., Г.ЮЖНО-САХАЛИНСК, ПР.МИРА 1А
ЮЖНО-САХАЛИНСК ПР.МИРА 1А
САХАЛИНСКАЯ ОБЛ., Г.ЮЖНО-САХАЛИНСК, ПР.МИРА Д.1А
ЮЖНО-САХАЛИНСК ПР.МИРА Д.1 СТР.А

Требуется преобразование значений полей адресов в структуру упоминания, содержащую, в частности, атрибуты: почтовый индекс, название страны, квалификатор региона, название региона, квалификатор района, название района; квалификатор города, название города; квалификатор городского района, название городского района; квалификатор подчинённого города, название подчинённого города; квалификатор улицы, название улицы, квалификатор номера дома; номер дома; квалификатор номера корпуса; номер корпуса, квалификатор номера квартиры; номер квартиры, нормализованная строка адреса. Атрибуты, заполняемые в соответствии со справочниками (квалификаторы и пр.), должны содержать вместо (помимо) значения, его код по соответствующему справочнику.

Такого рода семантическая нормализация является самостоятельной достаточно сложной задачей. Она может решаться специализированными процедурами, реализующими в своем коде логику отображения для различных реквизитов и их представлений, либо универсальными программными средствами, проводящими преобразования в соответствии с заданными моделями. В обоих случаях, для получения качественного результата важно использование эталонных справочников и классификаторов. Например, для нормализации российских адресов может использоваться справочник КЛАДР, ведение которого осуществляется налоговыми органами.
Более подробное описание подходов и методов, применяемых в контексте программных средств семантической нормализации, выходит за рамки данной статьи. Скажем только, что в разных проектах эта задача решалась разными средствами и в разном объеме.

Помимо преобразований, учитывающих семантику атрибутов, для всех полей определенных типов должны быть применены стандартные форматные преобразования, предназначенные для унификации представления их значений, например: приведение строк к одному регистру, отсечение лидирующих и финальных пробелов, замена табуляций и множественных пробелов единичными пробелами, преобразование различных форматов дат к единому формату и т.д.

5. Идентификация

Следующим этапом при подключении к Картотеке сведений о новом субъекте является идентификация. Под идентификацией подразумевается процесс оценки степени совпадения объектов и принятия решения о соотнесении упоминаний карточкам субъектов. В технологическом цикле формирования Картотеки подразумевается проведение автоматической идентификации упоминания при его подключении к реестру (технологическая схема подключения информации о субъекте приведена на рис. 3).

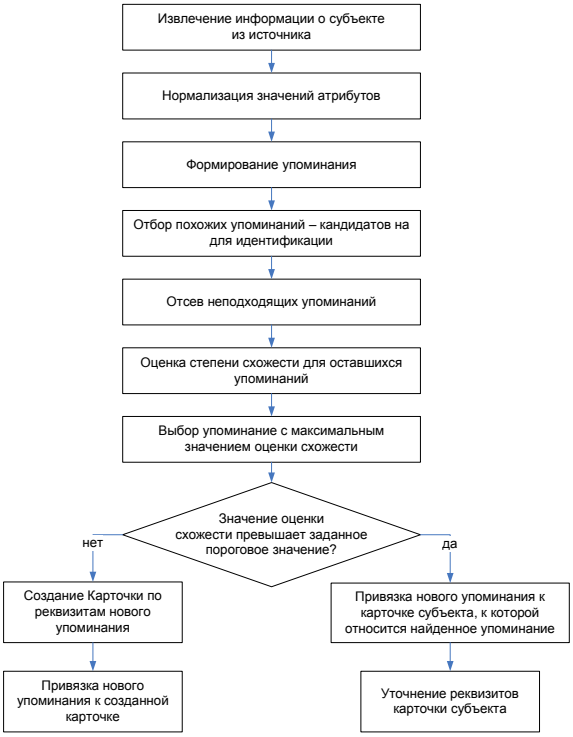

Рис. 3. Последовательность подключения сведений о субъекте.

Как видно на схеме, для нового упоминания последовательно исполняются шаги отбора, отсева и оценки схожести, поэтому предложенный метод условно назван методом трех «О». Правила, критерии и условия для каждого из шагов задаются в специализированной модели идентификации, которая индивидуально настраивается для каждого источника.
На первом шаге производится отбор упоминаний - кандидатов на идентификацию. Этот отбор

должен сформировать множество рассматриваемых упоминаний (из практики – до нескольких сотен) для дальнейшей обработки. Именно на этом шаге новое упоминание сопоставляется всему массиву упоминаний Картотеки, объем которого может составлять сотни миллионов записей. Поэтому критерии отбора в большинстве случаев строятся на точном совпадении отдельных реквизитов и использовании индексов с высокой степенью селективности. Для эффективного решения задачи отбора в схеме данных Картотеки максимально задействуются возможности СУБД: секционирование таблиц и индексов, использование битовых индексов и пр.

На втором шаге производится отсев тех из выбранных упоминаний, которые можно заведомо оценить как неподходящие. Как правило, критерием отсева является несовпадение значений ключевых реквизитов. Смысл этого этапа – максимально сократить множество упоминаний, задействованное на стадии оценки.

На третьем шаге производится более детальное сопоставление нового упоминания с оставшимися упоминаниями – кандидатами на идентификацию и вычисление оценки степени схожести. Расчет производится на основании заданных в модели идентификации весовых коэффициентов по отдельным атрибутам упоминания. Оценка степени схожести в простейшем случае получается суммированием коэффициентов по совпавшим атрибутам. В более развитом случае в ее расчете может учитываться степень похожести значений строковых атрибутов.

Задача оценки схожести может решаться специализированными процедурами, реализующими в своем коде логику сравнения структур упоминаний, либо универсальными программными средствами, производящими подобные расчеты для произвольных входных данных в соответствии с заданными моделями. В любом случае, этап оценки является достаточно дорогим с точки зрения использования вычислительных ресурсов.

Пример простейшей модели идентификации, приведенный ниже, иллюстрирует принципы, положенные в основу автоматической идентификации в Картотеке:

1. Критерии отбора (обрабатываются последовательно):
- совпадение ИНН;
- совпадение ОГРН;
- совпадение ОКПО;
2. Критерии отсева (отрабатывают только в том случае, если реквизит в обеих записях имеет непустое значение):
- несовпадение ИНН;
- несовпадение ОГРН;
- несовпадение организационно-правовой формы;
- несовпадение региона;
3. Коэффициенты оценки схожести:
- совпадение ИНН - 40;
- совпадение ОГРН - 50;
- совпадение краткого наименования - 30;
- совпадение нормализованной адресной строки - 20;
4. Пороговое значение идентификации - 100:

Такого рода модель (в реальности существенно более сложная) может быть сохранена в репозитории системы и использована при работе процедур обработки данных.

Изложенный метод автоматической идентификации, с одной стороны, обеспечивает достаточную гибкость системы, а с другой – ее необходимую производительность, так как во многих случаях существуют достаточно жесткие требования по скорости обработки больших информационных массивов.

Следует отметить, что вне зависимости от применяемых методов и технологий, задачи нормализации и идентификации данных о лицах, упоминаемых в различных федеральных и ведомственных базах данных, не могут автоматически решаться со 100% достоверностью результата. В любом случае, процесс наполнения и актуализации реестров субъектов требует участия эксперта качества данных, или информационного технолога, осуществляющего контроль, а в отдельных случаях и корректировку результатов автоматической обработки данных.

6. Пример проекта

Заказчик - Федеральная Таможенная Служба, проект - разработка Центрального реестра субъектов внешнеэкономической деятельности (ЦРСВЭД) - 2004 г.

Задача: сбор, идентификация и хранение согласованных сведений о юридических лицах, когда-либо вступавших в правоотношения с таможенными органами Российской Федерации (субъектах ВЭД) и имеющих упоминания как во внутренних информационных ресурсах (информационных системах таможенных органов), так и во внешних информационных ресурсах, полученных по каналам межведомственного обмена. Предоставление подразделениям таможенных органов объективной разносторонней информации о деятельности субъектов ВЭД.

Технологическая специфика. С одной стороны – это хранилище данных, интегрирующее информацию о юридических лицах, с другой стороны – совокупность учетных задач ведения реестров лиц, осуществляющих деятельность в области таможенного дела (таможенные брокеры, перевозчики и т.д.). Проведена интеграция более

10 разнородных источников, реализована репликация данных между узлами распределенной системы.
В данном проекте, как и в других проектах, реализованных на основе ВЕРО, применены основные положения обозначенного подхода. При этом программная реализация Картотеки в каждом из проектов имеет свои особенности в соответствии с предъявляемыми требованиями.

7. Границы применения

Подход к консолидации, основанный на ETL, обеспечивая высокое качество результата, предъявляет серьезные требования к вычислительным мощностям и дисковому пространству. Именно этими ресурсами системы оплачивается качество данных. Однако, как и любые ресурсы, они не беспредельны. А количество и объемы информационных источников потенциально неограниченны. Многие из этих источников имеют настолько низкое качество данных, что сведения из них в массе своей не могут быть идентифицированы.
Таким образом, следует сформулировать принципы, исходя из которых можно оценить целесообразность подключения определенного информационного источника к Картотеке.
Например, состав информации из различных источников, загружаемой, обрабатываемой и хранящейся в БД, построенной по принципу Картотеки, можно определять на основе оценки следующих характеристик информационных ресурсов:
• актуальность и достоверность (информационный ресурс должен поддерживаться и актуализироваться соответствующим уполномоченным органом, и поступать на основании заключенного соглашения о межведомственном информационном взаимодействии и утвержденного регламента информационного обмена);
• приемлемое качество данных для их автоматической обработки (состав и качество заполнения ключевых реквизитов должны обеспечивать возможность их автоматического сопоставления со сведениями в сфере деятельности ведомства – владельца системы);
• целесообразность загрузки в БД системы определенных данных с учетом информационных потребностей пользователей, необходимого для обработки этих данных ресурса вычислительных мощностей, а также общих требований к оперативности обработки данных и их качеству.

Таким образом, реализуемый в Картотеке подход по пакетной интеграции данных (ETL) применим в тех случаях, когда требуется:
• включение в общую технологию автоматизированной обработки и анализа информации достаточно хорошо документированных и надежных данных;
• интеграция ключевых справочных данных;
• удаление дублирующихся данных;
• осуществление процессов проверки качества данных.
При этом необходимые данные извлекаются, преобразуются и загружаются в БД Картотеки в виде, пригодном для их использования.
Для прочих информационных ресурсов, тем или иным образом оказавшихся доступными для использования (далее по тексту - дополнительных информационных ресурсов), можно предложить использование подхода Enterprise Information Integration (EII). За счет описанных на мета-уровне правил/моделей интеграции, инструменты EII обеспечивают универсальный метод доступа (запросов) к данным и используют технологию «вытягивания» информации из источников (pull technology). Использование в качестве основы построения решений EII сервисно-ориентированной архитектуры (SOA) поддерживает интеграцию и получение информации из структурированных, реляционных систем, а также из неструктурированных, источников (Web-контент, документы).

8. Аналогичные решения

Задача централизованного управления ключевой информацией о клиентах (Master Data Management) через специализированный центр – Hub, решается практически всеми ведущими мировыми софтверными компаниями — Microsoft, IBM, Oracle и др.

В частности, в продукте Oracle Customer Data Hub осуществляется централизация данных путем построения единого центрального хранилища информации о клиентах. В основе модели Customer Data Hub лежит архитектура TCA (Trading Community Architecture), изначально разработанная в рамках ERP-системы Oracle E-Business Suite. В дальнейшем фирма Oracle расширила функционал управления клиентскими данными и сейчас пакет Customer Data Hub предлагается в качестве независимого от системы E-Business Suite решения. Customer Data Hub базируется на принципе «главного набора данных» о клиенте и реализует идею «главного идентификатора» клиента, который может быть использован во всех приложениях и информационных системах предприятия.
Решение состоит из 3-х компонент:
1. Customer Data Model предоставляет возможность использования модели данных TCA, а также инфраструктуры TCA (включая открытый PL/SQL API, XML Web Services и пр.) для разработки собственных решений.

2. Oracle Customers Online представляет собой WEB-приложение для просмотра данных о клиентах, загрузки информации о клиентах из других систем, управления метаданными.
3. Oracle Data Librarian включает инструменты для сопоставления записей на основании системы правил соответствия, а также средства «слияния», сочетающие элементы ручной и автоматической обработки, что позволяет формировать «главный набор данных» о клиенте.

При загрузке данных из систем-источников, в БД Customer Data Hub сохраняется исходный идентификатор клиента. Это позволяет доработать информационные системы организации для реализации «обратной связи» с использованием предоставляемых Customer Data Hub программных интерфейсов. По сути, Oracle Customer Data Hub представляет собой программное ядро для консолидации корпоративных данных о клиентах, которое требует настройки на прикладную область Заказчика и разработки компонент интеграции с внешними системами.

Применяемые подходы в целом соответствуют изложенным в данной статье: преобразование данных при их загрузке из систем-источников, с последующей их идентификацией в соответствии с заданной моделью.

9. Заключение

Основные преимущества изложенного подхода: унификация структур хранения информации, процедур нормализации и идентификации, позволяющая обеспечить приемлемую производительность подключения и поиска на больших объемах данных и предоставить пользователю удобный и эффективный инструмент для поисковой и аналитической работы. Эти подходы были успешно применены (полностью или частично) в реальных программных проектах для ряда федеральных ведомств. Использование данной технологии целесообразно в случае разработки информационных систем, в которых объединяется информация о физических и юридических лицах из различных структурированных источников.

Литература

[1] Colin White. Master Data Management and Customer Data Integration. http://www.b-eye-network.com/view/2432

[2] А. Барченков, Т. Лякишева, А. Шаркин. Ставка на ВЕРО. Открытые системы, # 05-06/2005

[3] В. Васильев. ФОРС предлагает ВЕРО. Сетевой, #02/2005

[4] Леонид Черняк. Data Hub и MegaGrid. Computerworld #04/2005

[5] Jill Dyche, Evan Levy. Customer Data Integration: Reaching a Single Version of the Truth. // John Wiley & Sons Inc., 2006

Person Data Integration

Tatyana Lyakisheva
FORS Development Center
Tlyakish@fors.ru

This work focuses on problems of building master person indexes based on data about people and legal entities dispersed across multiple application systems and databases. The method of such indexes implementation is offered in conjunction with techniques of data consolidation and integration. The approaches discussed in this paper were applied in several custom development software projects.