РАЗРАБОТКА СИСТЕМ УПРАВЛЕНИЯ НОРМАТИВНО СПРАВОЧНОЙ ИНФОРМАЦИЕЙ ДЛЯ СИСТЕМ ОБРАБОТКИ СТАТИСТИЧЕСКОЙ ИНФОРМАЦИИ

К.А. Линев,
аспирант кафедры кибернетики Московского института электроники и математики (технического университета).
Адрес: г. Москва, Б.Трехсвятительский переулок, д. 3,
e-mail: ZnTenshi@hotmail.com.

В статье рассматривается задача построения систем управления нормативно справочной информацией (СУ НСИ) для систем сбора, контроля качества и обработки статистической информации. Выделяются особенности, характерные для СУ НСИ, предназначенных для использования в области автоматизации статистических исследований, формулируются требования к таким системам. Рассматриваются подходы к построению некоторых из компонентов изучаемых систем на основании опыта построения СУ НСИ для Всероссийской переписи населения 2010 года (ВПН-2010).

Ключевые слова: мастер-данные, древовидный спрвочник, стастититчкие обзоры, компьютерная поддержка, статитический справочник.

Под термином нормативно-справочная информация (НСИ), или мастер-данные, как правило, понимают условно-постоянную часть всей корпоративной (учрежденческой) информации, не претерпевающую существенных изменений в процессе повседневной деятельности организации, на основании которой формируются текущие документы. [1]

В крупных, особенно – территориально-распределенных компаниях, в силу исторического их развития, часто сосуществует большое количество действующих систем ведения НСИ, в том числе и не автоматизированных, а также различных справочников, часто никогда не предназначавшихся для использования в IT инфраструктуре. При этом в каждой такой системе присутствуют собственные источники пополнения НСИ. Эта ситуация оказывается серьезным препятствием на пути интеграции корпоративной IT-инфраструктуры и вызывает огромные трудности при обмене данными между локальными приложениями, а также при создании сводных аналитических отчетов.

Одним из ярких примеров предприятий, которым приходится работать с большим количеством НСИ, являются органы государственной статистики и частные статистические компании. Огромное

количество различных справочников и таблиц используется при обработке данных различных исследований, проводимых этими организациями. При этом справочники постоянно обновляются по результатам уже проведенных исследований, территориально-административных изменений, изменений законодательства и по множеству других причин. Рассмотрению вопросов организации СУ НСИ в контексте работы именно таких предприятий и посвящена эта работа.

СУ НСИ наряду с самой информацией включает также комплекс средств ее поиска, хранения, обработки и распределения, методов ее ведения, поддержания в актуальном состоянии, а также совокупность организационно-распорядительных документов и регламентов, регулирующих использование и ведение данных НСИ. [2]

Любая претендующая на промышленное использование информационная система должна поддерживать управляемые ею данные на высоком уровне качества. Важную роль играют критерии, которые на сегодня универсальны для любых типов корпоративных данных, такие как полнота, непротиворечивость, корректность и актуальность. Причем применительно к данным НСИ, жизненный цикл которых по определению превышает аналогичный цикл для оперативных данных, они имеют еще большее значение.

Вместе с тем, помимо этих классических критериев (реализация которых на сегодня обеспечивается вполне отработанными методиками проектирования данных и надежными программными продуктами), существуют и более специфические, характерные именно для НСИ. Это идентифицируемость и уникальность, которые обеспечивают однозначную и уникальную идентификацию данных, что необходимо для установления ссылок на них из других элементов НСИ и прикладных документов [2].

СУ НСИ для систем информационного обеспечения статистических исследований обладают рядом особенностей, которые приводят к особенной актуальности качественного управления НСИ в таких системах. Перечислим наиболее заметные из этих особенностей.

Полный цикл жизни СУ НСИ в течение сравнительно небольшого промежутка времени. Поскольку практически каждое статистическое исследование уникально за счет как различия требований разных организаций к содержимому статистической информации, так и изменения представления о содержании исследования на основании уже проведенных аналогичных исследований, для каждого исследования необходимо заново производить первичную загрузку и развертывание системы. Каждое исследование также потребует новых способов работы с бизнес-приложениями.

Высокая стоимость восстановления утерянных в результате ошибок данных. Поскольку сбор информации и представляет собой суть исследования, в случае ее потери стоимость восстановления утраченных данных равняется стоимости их изначального получения. Таким образом, каждый сбой в СУ НСИ наносит серьезный урон всему исследованию. Это в свою очередь делает низкокачественное управление НСИ неприемлемым в информационных системах, связанных с обработкой статистической информации.

Географическая удаленность элементов системы друг от друга. Любое крупное исследование связано с управлением большими объемами данных, получаемыми в удаленных друг от друга узлах системы. Это создает ряд специфических сложностей, начиная с проблем со связью между узлами, заканчивая различиями в местном времени.

Централизованность. В силу природы поставленной задачи, поток информации при сборе статистической информации всегда направлен к одному центральному узлу системы, в котором должна быть произведена обработка собранных данных. Эту особенность таких систем можно использовать при построении СУ НСИ для упрощения, а соответственно – повышения надежности таких систем.

Таким образом, разработка программного обеспечения для СУ НСИ сталкивается со следующим набором задач:

- ✦ Первичная загрузка мастер-данных
- ✦ Организация хранения мастер-данных
- ✦ Организация обновления и распространения мастер-данных
- ✦ Организация интерфейсов с бизнес-приложениями.

При выполнении всех этих операций, НСИ должна в каждый момент времени отвечать обозначенному выше набору критериев качества.

Для каждой из этих задач характерен собственный круг вопросов, которые необходимо решить при разработке ПО СУ НСИ.

При первичной загрузке данных возникает задача преобразования большого объема существующих разнородных справочников к некоторому эталонному виду. Типичной является ситуация, когда при проведении очередного большого статистического исследования приходится создавать для обработ-

ки его результатов отдельную информационную систему. Наиболее характерный способ хранения справочников в компаниях со слабой автоматизацией бизнес-процессов – электронные таблицы. Как правило, это файлы Excel или легко приводимые к ним форматы. Конечно, можно вручную переносить справочники, однако объемы справочной информации могут быть невероятно большими. Например, представьте себе всероссийский справочник, сопоставляющий городские кварталы и сельские населенные пункты их индексам. Даже если проделать всю эту работу вручную, количество допущенных в ней ошибок может обесценить справочник. Таким образом, возникает вопрос о необходимости автоматического преобразования данных.

Для решения этой проблемы в рамках проекта по техническому обеспечению Всероссийской переписи населения 2010 года была решена задача автоматического преобразования множества взаимосвязанных свободно редактируемых справочников, сохраненных в формате Excel в инструкции на языке T-SQL для СУБД Microcoft SQL Server по заполнению таблиц БД содержащимися в справочниках мастер-данными. Поскольку таблицы Word и многих других офисных приложений легко преобразуются в таблицы Exel путем прямого копирования данных, это решение фактически позволяет решить проблему первичного заполнения БД, входящей в СУ НСИ на основании эталонных справочников, хранящихся в виде электронных документов.

Приложение использует при работе метаданные о структуре справочников в формате XML и с помощью компонента Aspose Cells производит обработку документов Excel. Оно поддерживает сложные представления данных, такие, как древовидные справочники или связь записей различных справочников на основании их геометрического расположения на листе Excel. При загрузке производится контроль целостности данных, то есть загруженные мастер-данные гарантированно имеют корректную структуру.

Вторым по популярности после электронных документов способом хранения справочников в организациях являются «малые» СУБД, такие, как, например, Microsoft Access. Причем зачастую БД, в которых хранится НСИ, спроектирована неудачно и не обеспечивает должного качества НСИ. Кроме того, необходимо обеспечение связи между справочниками, загружаемыми из различных, до того не связанных источников.

В рамках того же проекта было создано приложение, решающее и эту задачу. Приложение работало с объемным, порядка 180000 записей, территориальным справочником, имеющим древовидную структуру. Оно обеспечивало обработку множественных локальных БД, содержащих часть этого справочника в виде плоских таблиц, и заполняло централизованный справочник с обеспечением всех нужных ссылок для формирования древовидного справочника.

После загрузки данных необходимо решить задачу об организации их хранения, обновления и распространения. В этом вопросе существуют три возможных подхода [1]:

- ✦ Централизованный, характеризующийся централизованным хранением эталонов мастер-данных
- ✦ Децентрализованный, характеризующийся созданием распределенного виртуального хранилища НСИ
- ✦ Смешанный, представляющий собой попытку объединить наилучшие качества предыдущих двух подходов.

Первый поход обладает целым рядом преимуществ, таких, как простота разработки и автоматическое решение проблемы поддержания целостности данных при условии замены всех копий на эталоны.

С другой стороны, представим себе систему, узлы которой расположены по всей России, каковым свойством обладает система, обеспечивающая работу любого всероссийского статистического исследования. Необходимость при каждой операции обращаться к некоторому центральному хранилищу, допустим, расположенному в Москве, не только приведет к катастрофическому падению скорости работы системы в узлах, расположенных в восточной Сибири, но и в ряде случаев просто сделает работу системы невозможной из-за отсутствия каналов связи с центральной системой.

Однако и традиционный распределенный подход в случае с всероссийским исследованием также не годится. Традиционным методом обновления справочников и поддержания целостности системы является ночная нормализация данных, когда никакие пользовательские операции не производятся. Однако что будет делать такая система, если в одном ее узле полночь, а в другом скоро полдень?

Таким образом, необходимо разработать некоторый гибридный механизм работы с распределенными данными, позволяющий обеспечить максимум преимуществ централизованного подхода, но при этом способный справиться со специфическими трудностями при работе на обширной территории.

Для решения этой проблемы предлагается с помощью методов теории графов оптимальным с точки зрения затрат сетевого трафика и времени образом разделить сеть на подсети, управляемые централизованным способом. Каждая такая подсеть должна быть связана с другими подсетями с помощью специального набора программных инструментов, позволяющего незаметно для подсети обеспечивать доставку в нее обновлений из центрального узла и передачу результатов работы. Важно отметить, что для обеспечения максимальной эффективности должны поддерживаться самые различные способы связи, начиная от защищенной передачи через Интернет и заканчивая транспортировкой данных на физических носителях.

В связи с последним требованием приходиться признать, что информационная система, скорее всего, сможет гарантировать актуальность данных только в центральном узле, в подсетях же будет поддерживаться только локальная целостность и информация, необходимая для связывания данных с, возможно, изменившимися данными центрального узла в тот момент, когда это будет возможно.

Таким образом, будет обеспечена, с одной стороны, централизация управления НСИ в каждой отдельной подсети, с другой стороны, ценой необходимости поддержания репликации данных лишь на небольшом числе узлов будет достигнута гибкость и масштабируемость распределенной системы.

На данном этапе в ходе работы над СУ НСИ для статистических исследований удалось выделить свойственные таким системам особенности, которые должны учитываться при разработке. Поставлены основные задачи, которые должны решаться системой. Продемонстрирована возможность эффективного решения одной из поставленных задач – выполнения первичной загрузки данных из имеющейся у клиентской организации слабоструктурированной информации – путем создания модулей, выполняющих загрузку данных из электронных документов и слабо структурированных БД в систему таблиц НСИ заданной структуры. Предложен способ решения задачи организации хранения, распространения и обновления данных с помощью двухуровневой схемы хранения, сочетающей свойства централизованной и децентрализованной моделей хранения данных.

В настоящее время производится исследование возможностей по практической реализации предложенной в статье схемы хранения, обновления и распространения данных. Кроме этого ведется разработка механизма взаимодействия СУ НСИ с бизнес-приложениями, который предоставлял бы возможности максимально гибкого ввода и вывода данных в различных представлениях, сохраняя при этом идентифицируемость и уникальность всех записей, управляемых системой. ■

Литература

1. Ярослав Помазков, «Системы НСИ: мировой опыт и тенденции развития», журнал PC Week, №522, Москва, 2006 г.
2. Дмитрий Гулько, «Мастер-данные: найден кратчайший путь к СОА», CNews, Москва, 2006 г.