МЕТОДЫ И ТЕХНОЛОГИИ ПОСТРОЕНИЯ ХРАНИЛИЩА ДАННЫХ И ЗНАНИЙ ДЛЯ ИССЛЕДОВАНИЙ ЭНЕРГЕТИКИ

А.Н. Копайгородский, Л.В. Массель

Введение. В Институте систем энергетики им. Л.А. Мелентьева СО РАН выполняются исследования систем энергетики (электроэнергетики, тепло-, газо-, угле-, нефте-, нефтепродуктоснабжения), исследования энергетической безопасности России, региональных проблем энергетики, взаимосвязей энергетики и экономики, работы выполняются для стран СНГ, России и ее регионов. В рамках основных научных направлений выполняются исследования развития и функционирования как отраслевых систем энергетики, так и топливно-энергетического комплекса в целом. Результаты исследований отраслевых систем энергетики зачастую являются исходными данными для исследований ТЭК, а результаты исследований направлений развития ТЭК должны учитываться при исследованиях развития отраслевых систем энергетики [1].

В информационном обеспечении исследований энергетики можно выделить две взаимосвязанные, но в то же время различные проблемы:

1. проблема информационной обеспеченности, т.е. обеспеченности данными, связанная с необходимостью получения данных из разных источников, их верификации (оценка достоверности как источников, так и самих данных, устранение ошибок и разночтений и т.д.);
2. проблема разработки инструментальных средств информационного обеспечения.

Первая проблема связана с затрудненностью получения необходимых данных и является в ряде случаев серьезным препятствием для научных исследований. Инструментальные средства информационного обеспечения исследований энергетики эволюционировали параллельно с программным обеспечением, так же в институте ведутся работы по созданию качественно новых инструментальных средств.

В настоящее время, когда технические проблемы решены, на первый план вышла проблема информационной обеспеченности, так как практически невозможно получение данных с одинаковой степенью детальности по всем отраслевым системам энергетики. С учетом появления современных технологий хранения данных, ориентированных на корпоративное использование [2], реализуется общая информационная база в виде корпоративного хранилища данных - Репозитария ИТ-инфраструктуры [3-5], на качественно иной основе интегрирующего операционные (использующиеся для расчетов) базы данных, имеющиеся в ИСЭМ СО РАН.

Однако до сих пор остается не решенной проблема поддержки исследований отдельных отраслевых систем ТЭК: выполняя анализ существующих проблем и занимаясь прогнозированием развития систем энергетики, исследователю приходится обрабатывать огромный массив данных с помощью типовых либо специализированных программных средств. Исходные данные для выполнения работ исследователи получают из различных источников, данные могут представляться в различных форматах.

Предлагаемый подход. Для поддержки исследований отдельных систем энергетики авторами предлагается использовать специализированные хранилища данных и хранилища знаний для каждой системы энергетики. Под хранилищем данных понимается предметно-ориентированный, интегрированный неизменяемый набор данных с поддержкой хронологии записи данных, необходимый для принятия решений [2]. Для систематизации и накопления знаний о предметной области, представленных в виде документов (статей, отчетов и др.), используется хранилище знаний. Под знаниями о предметной области в первую очередь понимаются декларативные явные знания [6], но в системе также предусмотрена возможность хранения процедурных знаний (описания программ и алгоритмов). Метаданные позволяют описывать знания, выполнять их классификацию и каталогизацию, и используются для быстрого и удобного поиска. При применении "типовых решений" поддержки исследований отдельных систем задача построения единого корпоративного хранилища для решения комплексных проблем энергетики значительно упрощается.

В ИСЭМ СО РАН на протяжении ряда лет ведутся работы по созданию ИТ-инфраструктуры исследований энергетики [3,4], которая призвана облегчить разработку и использование различных информационных и вычислительных ресурсов. ИТ-инфраструктура состоит из четырех основных составляющих: интеллектуальной, информационной, вычислительной и телекоммуникационной инфраструктуры.

Информационная инфраструктура [5] объединяет информацию обо всех разрозненных базах данных, программных комплексах моделях данных, моделях программ, представленных в виде UML, ERD, XML и др. Программные компоненты информационной инфраструктуры создаются на основе концепции сервис-ориентированной архитектуры (SOA): с одной стороны, компоненты обеспечивают выполнение достаточно простых функций, с другой, применение компонентов в определенной последовательности позволяет решать достаточно сложные задачи. Применение готовых компонентов позволяет ускорить реализацию хранилища данных и знаний для поддержки исследований систем энергетики.

Архитектура хранилища данных и знаний. Процесс исследования любой энергетической системы начинается со сбора массива исходных данных, который может быть получен из различных статей, отчетов,

статистических сборников, также в качестве исходных данных могут выступать результаты предыдущих исследований. Внесение информации выполняется с привязкой к словарю предметной области: исследователь должен выполнить сопоставление определенных отчетных или статистических данных с регионом, категорией ресурса, его целевым назначением, должен указать и другие классификационные характеристики. В хранилище данных и хранилище знаний отдельной отрасли энергетики словарь предметной области является общим (одним) и содержит свойственные ей классификаторы (рис. 1). Метаданные также являются общими и описывают как структуру данных, так и документы, помещаемые в хранилище. Таким образом, хранилище данных и знаний состоит из четырех основных логических частей: словаря предметной области, метаданных, непосредственно данных хранилища, которые физически расположены в базе данных, и декларативных знаний, представленных в виде документов, которые находятся в файловом хранилище. Стоит отметить, что ограничения накладываются только на структуру метаданных, которые описывают документы, находящиеся в хранилище, модели словаря предметной области и хранилища данных.

Репозитарий является одним из основных компонентов ИТ-инфраструктуры исследований энергетики и содержит информацию обо всех других компонентах, их местоположении и о способах доступа к ним [5]. В нем описываются хранилища данных и знаний отдельных систем энергетики, указывается их расположение (адреса серверов) и интерфейсы взаимодействия (описания Web-сервисов). В Репозитарии также описаны оперативные базы данных, используемые в исследованиях, программные комплексы, научные труды сотрудников института и др.

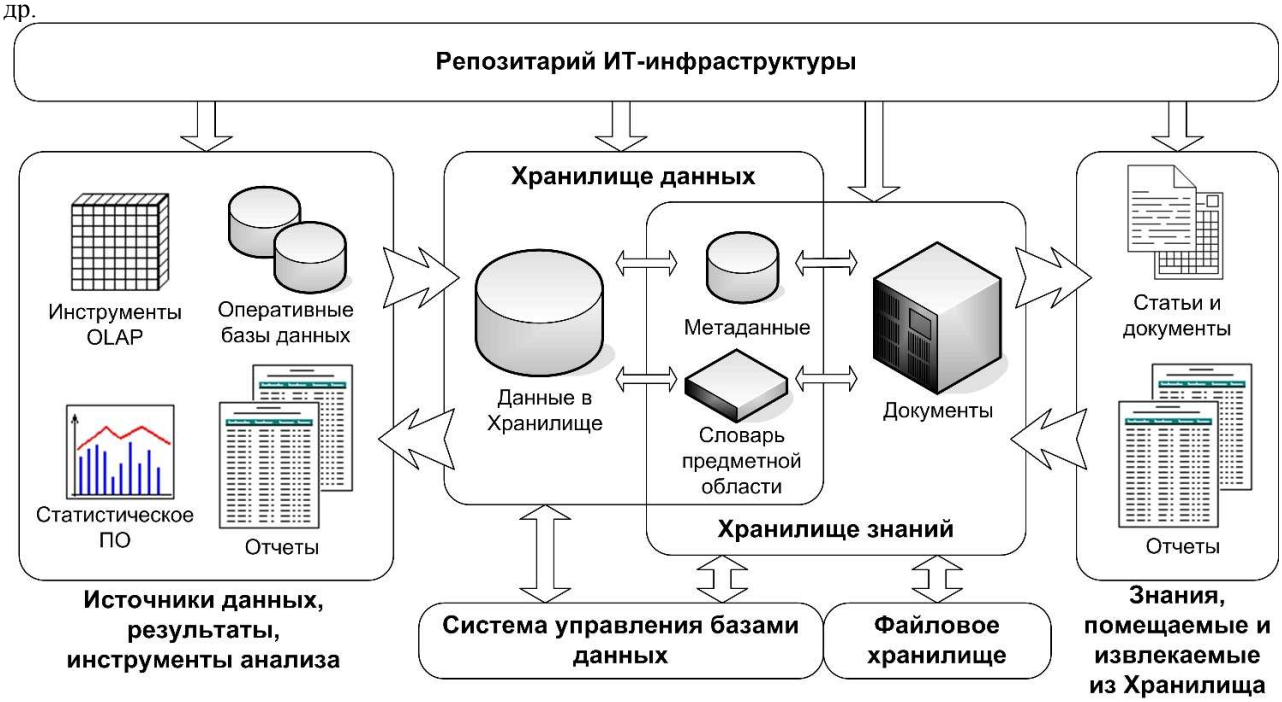

Рис. 1. Архитектура хранилища данных и знаний исследований энергетики

Построение хранилища знаний. Хранилище знаний предназначено для накопления и структурирования декларативных явных знаний предметной области. Основным понятием, которым оперирует хранилище, является "Документ". Документ имеет аннотацию, некоторую структуру (содержание, список таблиц, рисунков), содержит информацию об авторах, дате и месте публикации (в том числе организацию-издателя и URL электронного документа), связан с ключевыми словами и классификаторами из словаря предметной области. Кроме метаданных (описаний) документов в хранилище знаний могут содержаться и их полные тексты. Для более удобного представления документов и работы с ними в хранилище предусмотрена возможность создания логических группировок на основе метаданных - витрин документов. Под витриной документов понимается виртуальная совокупность документов хранилища, выделенная по какому-либо признаку или для определенной цели [7]. Например, в хранилище можно создать такие витрины документов как "статьи", "книги", "статистические сборники", "издания 2009 года по СФО", "подготовка отчета по НИР". Витрины документов формируются на основе поисковых запросов, и могут содержать как шаблоны запросов, которые будут всякий раз выполняться при выборе той или иной витрины, так и результаты запроса, которые являются статичными.

Для обеспечения безопасности знаний в хранилище предусмотрено использование симметричного шифрования. Алгоритмы шифрования могут быть применены только к полным текстам документов, расположенным в хранилище. Использование именно симметричного шифрования обусловлено тем, что шифрование и дешифровка данных выполняется на стороне клиента (на одном и том же компьютере), поэтому применение асимметричных алгоритмов не является целесообразным. Основными задачами файлового

хранилища являются размещение, передача файлов пользователю и их удаление, поэтому содержание файлов всегда находится в зашифрованном виде.

Для обеспечения многопользовательского доступа к шифрованным данным без размещения ключей на всех компьютерах может использоваться криптошлюз. Ключи шифрования размещаются на этом защищенном узле, с указанием пользователей и документов к которым они могут быть применены. При запросе данных через криптошлюз, он выполняет эквивалентный запрос к хранилищу данных, дешифрование и передает результат пользователю.

Построение хранилища данных. В процессе исследований функционирования и развития энергетических систем приходится оперировать достаточно большими объемами данных, получаемых из различных источников. Большой объем данных обусловлен их временным характером и множеством показателей исследуемых объектов энергетики. Информация размещается в хранилище данных в соответствии с созданной моделью для выбранной системы энергетики. После внесения данных исследователь имеет возможность выполнить их анализ, выгрузку в различные форматы, использовать полученные данные в качестве исходной информации для специализированных программ моделирования. Таким образом, основная сложность реализации хранилища данных для поддержки исследований систем энергетики заключается в создании достаточно универсальных механизмов импорта и экспорта данных, а также в описании модели предметной области внутри хранилища.

Задачи импорта и экспорта данных в хранилище выполняются в два этапа с применением промежуточного формата хранения подготовленных данных (Structured Data File - SDF). При загрузке данных на первом шаге они преобразуются в SDF-формат, а затем выполняется загрузка SDF-файлов в хранилище. При экспорте - данные извлекаются в промежуточном формате SDF, а затем, с помощью специализированных средств конвертирования, могут быть преобразованы в различные документы: RTF, TXT, DBF, HTML, Microsoft Word, Microsoft Excel и др. При этом не накладывается жестких ограничений ни на форматы исходных данных, загружаемых в хранилище, - они могут быть представлены в различных СУБД или документах; ни на форматы выходных документов. Если необходима поддержка нового формата - потребуется лишь реализовать конвертор, который преобразует данные из SDF-формата в требуемый формат документов.

Одним из основных принципов построения хранилища данных для поддержки исследований систем энергетики является использование единой структуры метаданных (части схемы базы данных). Метаданные хранилища описывают лежащую в его основе модель данных исследуемой системы энергетики, структуру словаря предметной области, содержат регламентированные запросы, а также другую дополнительную информацию, используемую для автоматизации работы. Таким образом, становится возможным построение универсальных программных компонентов, взаимодействующих с хранилищем, не зависящих от исследуемой системы энергетики.

Тексты и параметры регламентированных запросов могут быть размещены в таблице с именем RepositoryQuery в хранилище данных либо расположены в оперативных базах данных. Использование в оперативных базах представления с идентичным именем дает возможность разработчикам реализовать разграничение прав для внешних пользователей. Применение параметров в регламентированных запросах предоставляет пользователю возможность устанавливать различные критерии поиска и извлечения информации [8].

Инструментальные средства поддержки хранилища данных и знаний реализуются на объектно-ориентированном языке Java (Java Standard Edition) в среде NetBeans [9,10]. В настоящее время в качестве базовой СУБД используется Firebird. Стоит отметить, что при выбранном подходе к созданию инструментальных средств и использованию метаданных возможен достаточно легкий переход почти на любую другую СУБД.

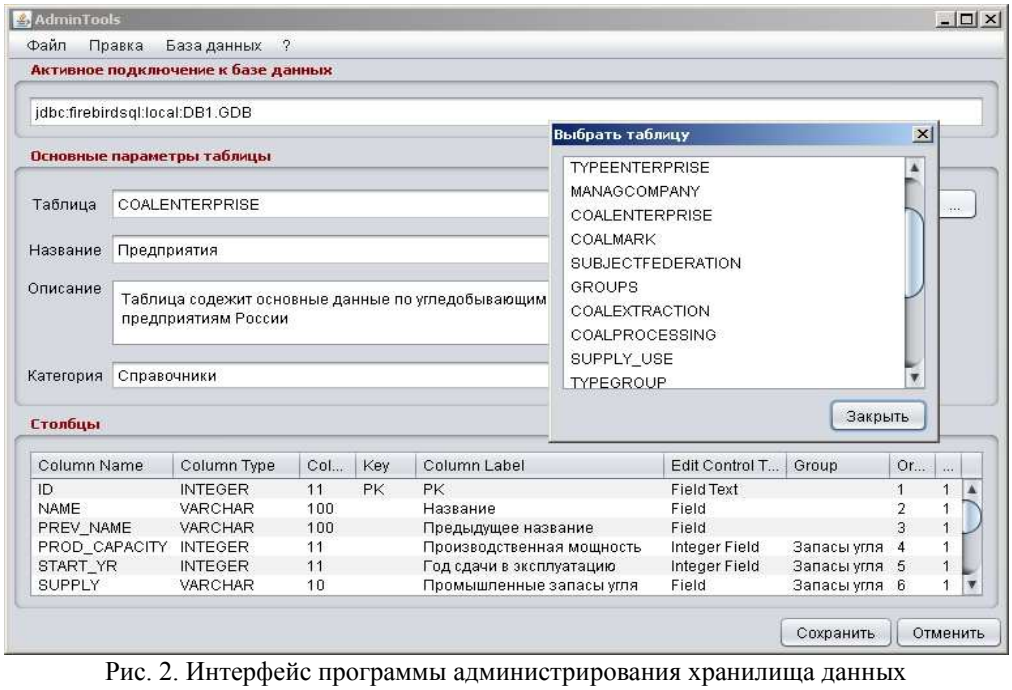

Рис. 2. Интерфейс программы администрирования хранилища данных

Программное обеспечение для работы с хранилищем данных и знаний состоит из отдельных компонентов (модулей):

1. программа администрирования хранилища данных (рис. 2) - предназначена для конфигурирования хранилища, описания модели данных и др.;
2. программа для работы с хранилищем данных (рис. 3) - применяется пользователями для просмотра, корректировки и извлечения данных, программа активно использует метаданные, расположенные в хранилище;
3. библиотека функций конвертации - реализует преобразования в формат SDF и из него, реализована в виде отдельного компонента, что позволяет легко дополнять ее и вносить изменения в существующие преобразования;
4. программа для работы с хранилищем знаний - позволяет вносить, описывать, находить и извлекать декларативные знания, представленные в виде документов.

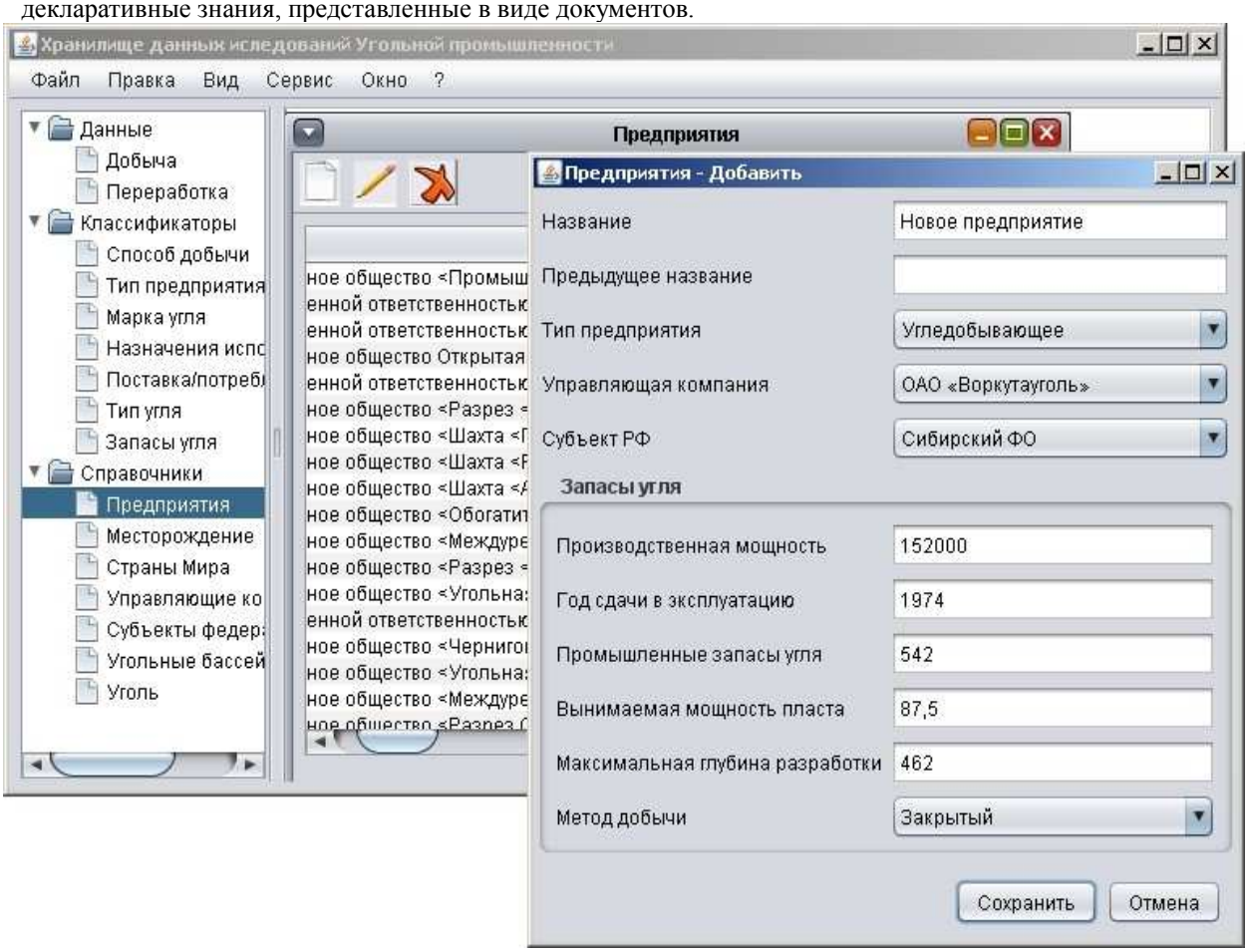

Рис. 3. Интерфейс программы для работы с хранилищем данных

Созданное хранилище данных и знаний интегрировано в ИТ-инфраструктуру исследований энергетики. Универсальные компоненты информационной инфраструктуры могут быть применены для решения различных задач: построения отчетов, преобразования данных для аналитических систем, извлечения и передачи данных в различные СУБД и загрузки данных из XML-файлов. Для этого используются генератор отчетов, программа извлечения данных и компонент загрузки структурированных данных. Генератор отчетов работает абсолютно прозрачно для других программ и никоим образом не влияет на структуру базы данных или на хранящиеся в ней данные. Программа извлечения данных поддерживает как регламентированные запросы, созданные администратором хранилища, так и произвольные пользовательские запросы, результаты их выполнения могут быть представлены в виде таблиц или OLAP-кубов.

Заключение. В статье описаны методы и технологии построения хранилища данных и знаний для исследования отраслевых систем ТЭК. Для поддержки этих исследований авторами предлагается использовать специализированное хранилище данных и знаний; метаданные, которые описывают исследуемую систему энергетики, структуру данных и документы, помещаемые в хранилище. Применение метаданных позволяет строить универсальные программные компоненты, взаимодействующие с хранилищем. Реализация инструментальных средств выполняется на объектно-ориентированном языке Java, в качестве базовой СУБД используется Firebird. Хранилище данных и знаний интегрировано в ИТ-инфраструктуру исследований энергетики, компоненты которой могут быть применены для решения различных задач. Исследования,

описанные в статье, выполнены при частичной финансовой поддержке грантов РФФИ №08-07-00172, №10-07-00264 и гранта Программы Президиума РАН №2.29.

ЛИТЕРАТУРА:

1. Беляев Л.С., Санеев Б.Г., Филиппов С.П. и др. Системные исследования проблем энергетики / под ред. Н.И. Воропая.- Новосибирск: Наука, 2000.- 558 с.
2. W. H. Inmon Building the Data Warehouse, Fourth Edition, 2005 Published by Wiley Published Publishing, Inc., Indianapolis, Indiana.
3. Воропай Н.И., Массель Л.В. ИТ-инфраструктура системных исследований в энергетике и предоставление ИТ-услуг. - Известия АН - Энергетика, №3, 2006.- С. 86-93.
4. Массель Л.В., Копайгородский А.Н. Технологии и система хранения данных и знаний для исследований в энергетике // Материалы Всероссийской конференции "Современные информационные технологии для научных исследований". Магадан: СВНЦ ДВО РАН, 2008.- С. 64-66.
5. Копайгородский А.Н., Массель Л.В. Разработка и интеграция основных компонентов информационной инфраструктуры научных исследований // Вестник ИрГТУ. - 2006. - № 2 (26).- С. 20-24.
6. Тузовский А.Ф., Чириков С.В., Ямпольский В.З. Системы управления знаниями (методы и технологии) / под ред. В.З. Ямпольского. - Томск: Изд-во НТЛ, 2005. - 260 с.
7. Такайшвили Л.Н., Осама Ель Сайед Шета. Проектирование хранилища документов для исследований развития угольной промышленности, Труды XIV Байкальской Всероссийской конференции "Информационные и математические технологии в науке и управлении". - Иркутск: ИСЭМ СО РАН, 2009. - С. 208-214.
8. Копайгородский А.Н. Виртуальная интеграция распределенных данных исследований в энергетике // Труды XIII Байкальской Всероссийской конференции "Информационные и математические технологии в науке и управлении". - Иркутск: ИСЭМ СО РАН, 2008. - С. 260-266.
9. Брюс Эккель. Философия Java (Thinking in Java).- 3-е изд.- СПб.: Питер, 2003. - 976 с.
10. Монахов В.В. Язык программирования Java и среда NetBeans.- 2-е изд.- СПб.: БХВ-Петербург, 2009.- 720 с.