электронное научно-техническое издание
НАУКА и ОБРАЗОВАНИЕ
Эл № ФС 77 - 30569. Государственная регистрация №0420900025. ISSN 1994-0408

Технология использования хранилища данных и знаний в исследованиях энергетики

10, октябрь 2010
автор: Осама Е. С.

oesheta75@gmail.com

Иркутский государственный технический университет

Введение. 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 [1]. Современный уровень развития информационных технологий позволяет формализовать и реализовать отдельные процедуры работы с документами. Для этих целей предлагается использовать специализированные хранилища данных и знаний для каждой системы энергетики. В статье рассматривается технология использования реализованного автором хранилища данных и знаний на примере исследований развития угольной промышленности.

Архитектура хранилища данных и знаний. Хранилище данных и знаний состоит из двух частей – хранилища данных и хранилища знаний (рис. 1).

Компоненты хранилища данных включают:

- программы администрирования хранилища данных – предназначены для конфигурирования хранилища, описания модели данных и др.;

- программы для работы с хранилищем данных – применяются пользователями для просмотра, корректировки и извлечения данных, программы активно используют метаданные, расположенные в хранилище;
- библиотека функций конвертации – реализует преобразование данных в различные документы (RTF, TXT, DBF, Microsoft Word, Microsoft Excel и др.) с использованием промежуточного формата SDF [2];
- ядро СУБД Firebird 1.5.2, организующее сетевой доступ к базе метаданных и обработку информации.

Компоненты хранилища знаний включают:

- программу для работы с хранилищем знаний – позволяет вносить, описывать, находить и извлекать декларативные знания, представленные в виде документов;
- ядро СУБД Firebird 1.5.2.

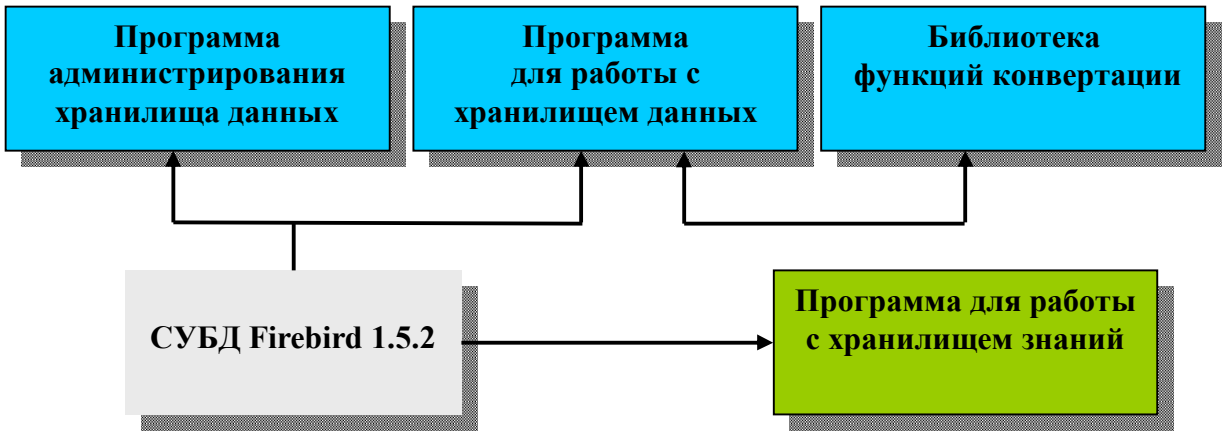


Рис. 1. Архитектура инструментальных средств поддержки хранилища данных и знаний

В качестве базовой СУБД был выбран сервер Firebird 1.5.2. Firebird является кроссплатформенной, компактной, свободно распространяемой СУБД. Основными преимуществами Firebird являются: клиент-серверная архитектура, обеспечивающая параллельную обработку запросов разных пользователей, высокая эффективность и мощная языковая поддержка в виде хранимых процедур и триггеров [3].

Технология использования хранилища данных. Технология использования хранилища данных включает несколько этапов: 1) построение модели данных предметной области – на этом этапе определяются правила хранения метаданных; 2) описание модели данных в виде таблиц данных – выполняется описание информационных ресурсов хранилища данных; 3) заполнение словаря предметной области – на этом этапе создаются словари предметных областей для просмотра, корректировки и извлечения данных, используются метаданные, помещенные в хранилище; 4) загрузка данных – использование хранилища данных для поддержки проведения исследований как интегрированного источника получения информации; 5) и 6) корректировка и выгрузка данных – поддерживаются процедуры ввода, вызова, изменения и экспорта данных. Перечень технологических этапов, инструментальных средств их поддержки и результатов этапа представлен в таблице 1.

Таблица 1. Технология использования хранилища данных и инструментальные средства.

Технологический этап	**Инструментальные средства поддержки**	**Результат**
1. Построение модели данных предметной области	Используется онтология, построенная в CmapTools* CASE-средство ERWin*	Модель данных, созданная на основе модели метаданных
2. Описание модели данных предметной области	Программа администрирования хранилища данных	Сконфигурированное хранилище данных
3. Заполнение словаря предметной области	Программа для работы с хранилищем данных	Созданный словарь предметной области для исследований энергетики (например, угольная

		промышленность)
4. Загрузка данных	Библиотека функций конвертации	Представление данных в различных форматах, передача в формат SDF
5. Корректировка данных	Программа для работы с хранилищем данных	Внесение данных в формат SDF для корректировки
6. Выгрузка данных	Библиотека функций конвертации	Представление данных в формате SDF , преобразование в различные форматы

*Информация о CASE-средствах дается ниже

С использованием инструментальной системы CmapTools было построено графическое представление модели данных предметной области в виде онтологии (рис. 2), которое описывает структуру Хранилища данных. Выбор CMAP Tools обусловлен тем, что она обладает рядом достоинств: позволяет создавать концепт-карты, добавлять ссылки на другие карты; имеет хороший и быстрый в освоении интерфейс; успешно применяется для создания достаточно крупных карт понятий; является бесплатной [4].

Из рисунка видно, что субъект федерации входит в группу субъектов, имеет месторождение и угольный бассейн. Каждый субъект федерации осуществляет добычу, поставку и потребление угля. Предприятие находится в субъекте федерации, управляется управляющей компанией, характеризуется типом предприятия и способом добычи. Также предприятие осуществляет добычу и переработку угля разных марок.

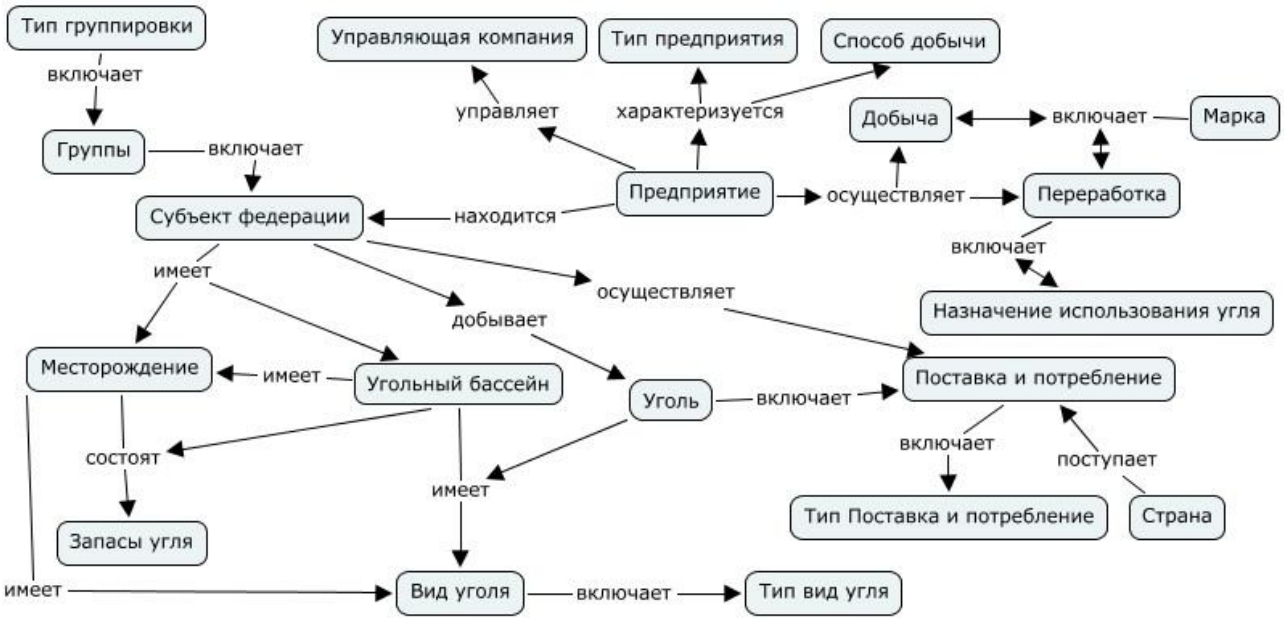


Рис. 2. Фрагмент онтологии предметной области (угольной промышленности)

На рис. 3 приведена модель данных, используемая для проектирования Хранилища данных угольной промышленности, построенная с помощью CASE-средства ERWin [5].

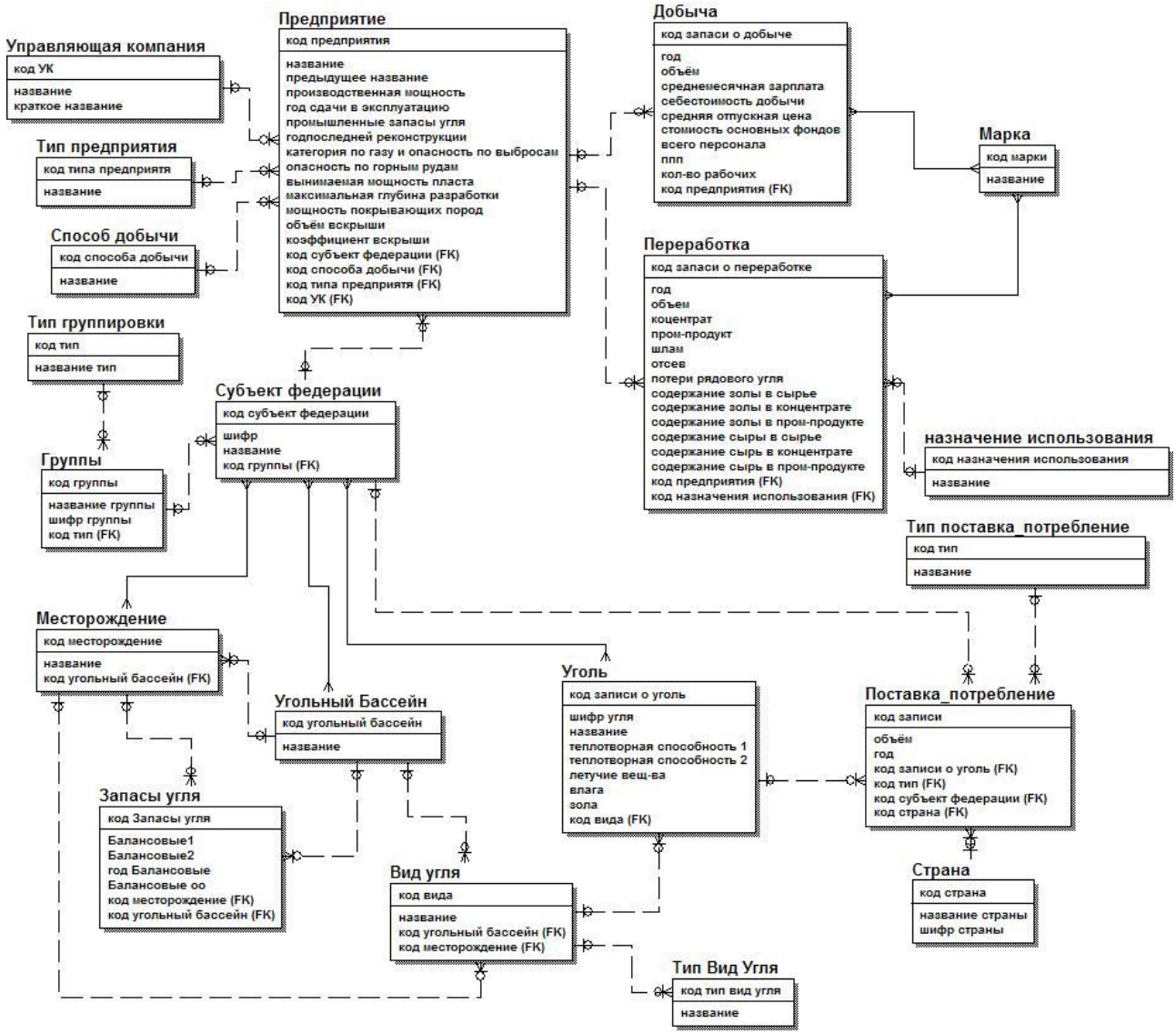


Рис. 3. Модель данных угольной промышленности в нотации ERWin

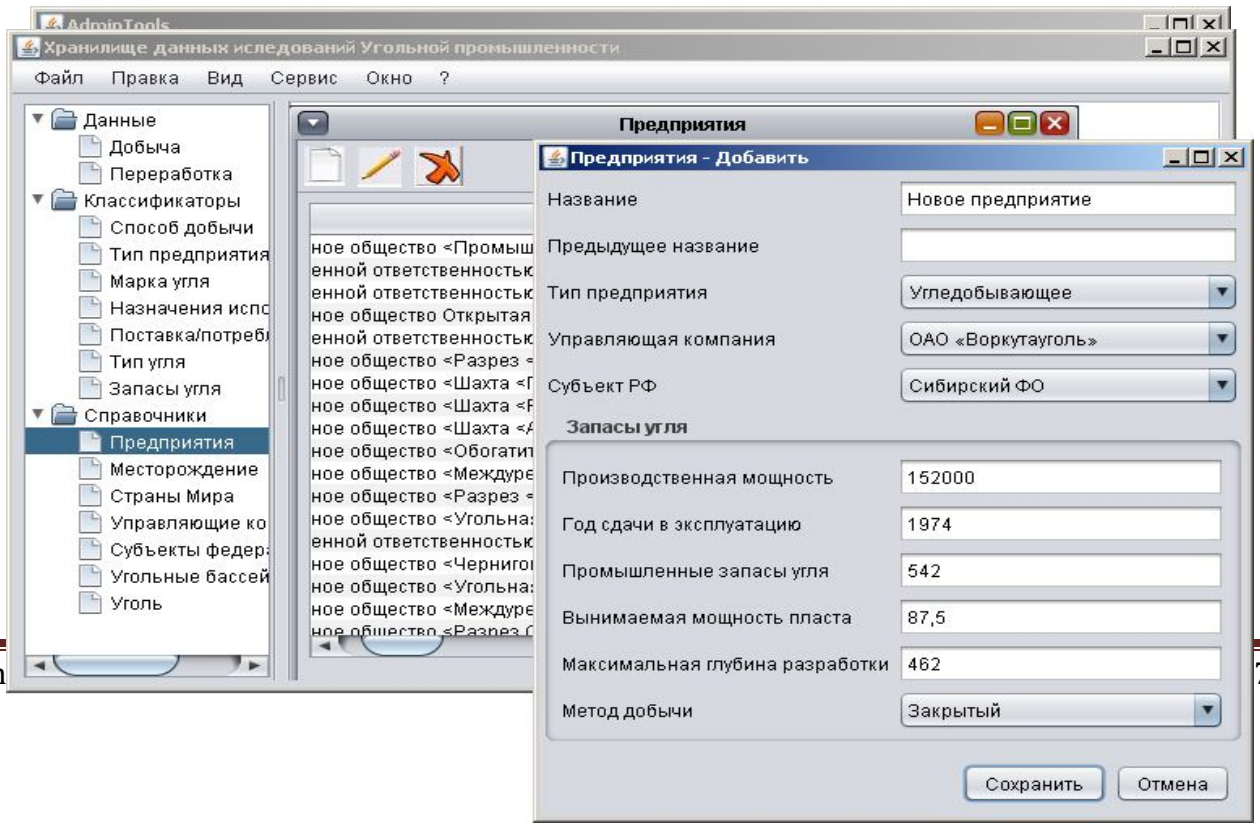
Программа администрирования хранилища данных, интерфейс которой представлен на рисунке 4, разработана специально для выполнения административных задач и подробно описана в работе [2]. Она позволяет выполнять модификацию структуры БД, поддерживать специальный набор данных, содержащих метаданные (сведения о структуре и семантике БД) и осуществлять ручное редактирование данных. Программа не привязана к конкретной предметной области и является инструментом, который существенно упрощает этап перевода знаний эксперта в набор таблиц реляционной базы данных и ее дальнейшее сопровождение. Визуальные средства отображают структуру БД в терминах объектно-ориентированного подхода (классы и свойства). Этот подход более естественен для человека

при описании сложных предметных областей, а возможности реляционных баз данных позволяют организовать эффективное хранение и обработку накапливаемой информации. Программа взаимодействует с СУБД через драйвер JDBC [6, 7], возможно подключать различные СУБД. Для заполнения словарей предметных областей в базах данных создано специальное приложение, которое базируется на метаданных (рис. 5). Основной целью при его разработке было построение такого приложения, которое давало бы возможность автоматически генерировать структуру словаря предметной области, формулировать регламентированные запросы, а также создавать другую дополнительную информацию, используемую для автоматизации работы. С помощью этого приложения пользователь извлекает необходимые данные из баз данных. Извлечение данных из СУБД Firebird 1.0/1.5 осуществляется через драйвер JDBC.

Рис. 4. Интерфейс программы администрирования хранилища данных

Рис. 5. Интерфейс программы для работы с хранилищем данных

Технология использования хранилища знаний. Технология использования хранилища знаний также разбивается на несколько этапов: 1) построение модели хранилища знаний



угольной промышленности – на этом этапе строится онтология предметной области; 2) заполнение словаря и конфигурирование хранилища знаний – на этом этапе создаются словари предметной области; 3) загрузка документов – документы помещаются в хранилище знаний; 4) формирование метаданных – описание документов и размещение их в витринах данных (витрина данных – часть хранилища данных, создаваемая для временного хранения документов одного типа); 5) поиск и извлечение документов – выполняются генерация временных данных и извлечение найденных документов. Перечень технологических этапов, инструментальных средств их поддержки и результатов представлен в таблице 2.

С использованием инструментальной системы CmapTools построено графическое представление модели хранилища знаний угольной промышленности в виде онтологии, описывающей информационные ресурсы хранилища знаний (рис. 6).

Таблица 2. Технология использования хранилища знаний и инструментальные средства.

Технологический этап	**Инструментальные средства поддержки**	**Результат**
1) Построение модели хранилища знаний угольной промышленности	Онтология CmapTools	Онтология - модель хранилища знаний угольной промышленности
2) Заполнение словарей и конфигурирование хранилища знаний	Программа администрирования хранилища знаний	Созданные словари предметной области
3) Загрузка документов	Программа администрирования хранилища знаний	Документы в хранилище знаний
4) Формирование метаданных	Программа администрирования хранилища знаний	Описание документов и размещение их в витринах

		данных
5) Поиск и извлечение документов	Программа администрирования хранилища знаний	Генерация временных данных и извлеченный документ.

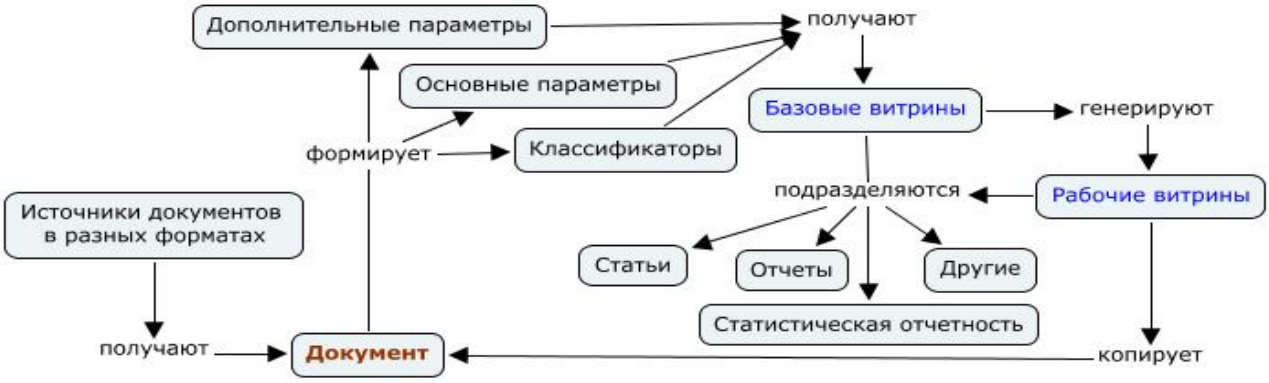


Рис. 6. Модель хранилища знаний угольной промышленности в виде онтологии

Технология использования Хранилища данных и знаний иллюстрируется рис. 7. Существуют источники документов, включающие все документы в разных форматах, которые хранились в базах данных систем энергетики или в другом электронном виде разными пользователями. Соответствующая процедура определяет тип документа.

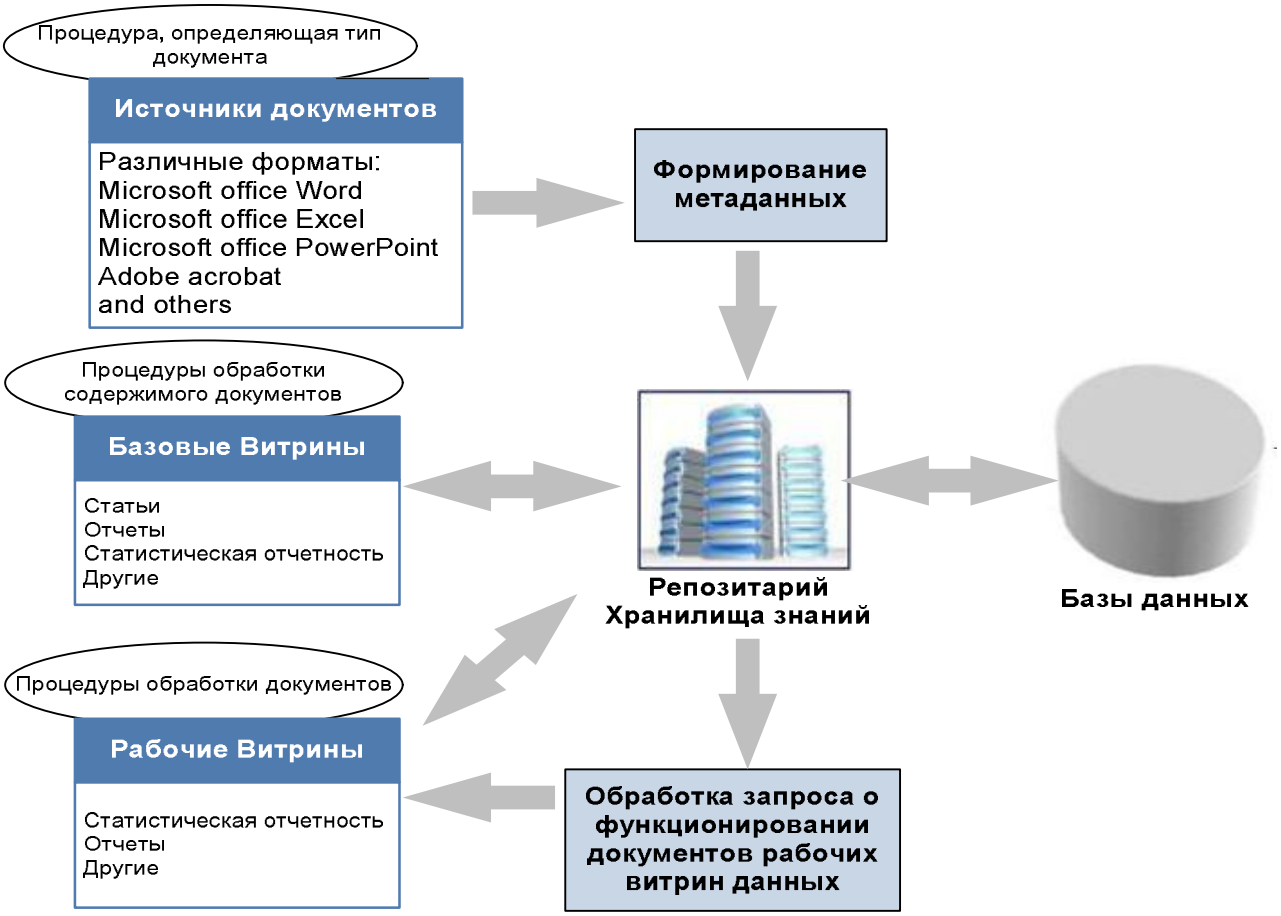


Рис. 7. Использование хранилища знаний угольной промышленности

Формирование метаданных включает в себя три этапа: формирование основных параметров, дополнительных параметров и классификаторов. Основные параметры включают в себя название документа, аннотацию, фамилии авторов, ключевые слова, комментарии, имя файла, источник документа, год, процент достоверности и комментарий. Дополнительные параметры включают: содержимое документа, таблицы, рисунки и приложения. Классификаторы необходимы для того, чтобы выбирать базовые витрины данных и типы классификаторов. Базовые витрины рассматриваются как создаваемые на длительное время виртуальные массивы документов, отобранных на основе некоторых параметров или со специальной целью и организованные в соответствии с потребностями конечных пользователей и спецификой предметной области.

Обработка запроса о функционировании документов рабочих витрин данных позволяет конечному пользователю или квалифицированному специалисту при проведении

исследований осуществлять поиск по таким поисковым признакам, как год, источники документов, классификаторы, авторы и ключевые слова (дескрипторы). Результатом поиска могут быть как ответы на запросы, так и рабочие (временные) витрины данных. Рабочие витрины создаются как временные виртуальные массивы данных, которые необходимы для определенной цели (написание статьи, подготовка отчета и т.д.); в дальнейшем специалист может использовать сгенерированные им рабочие витрины.

Программа администрирования хранилища знаний, интерфейс которой показан на рис. 8, подробно описана в работе [8]. Она разработана специально для исследователей, владельцев информации и пользователей, которые, кроме традиционных способов, будут фиксировать и сохранять свои знания в структурированном электронном виде.

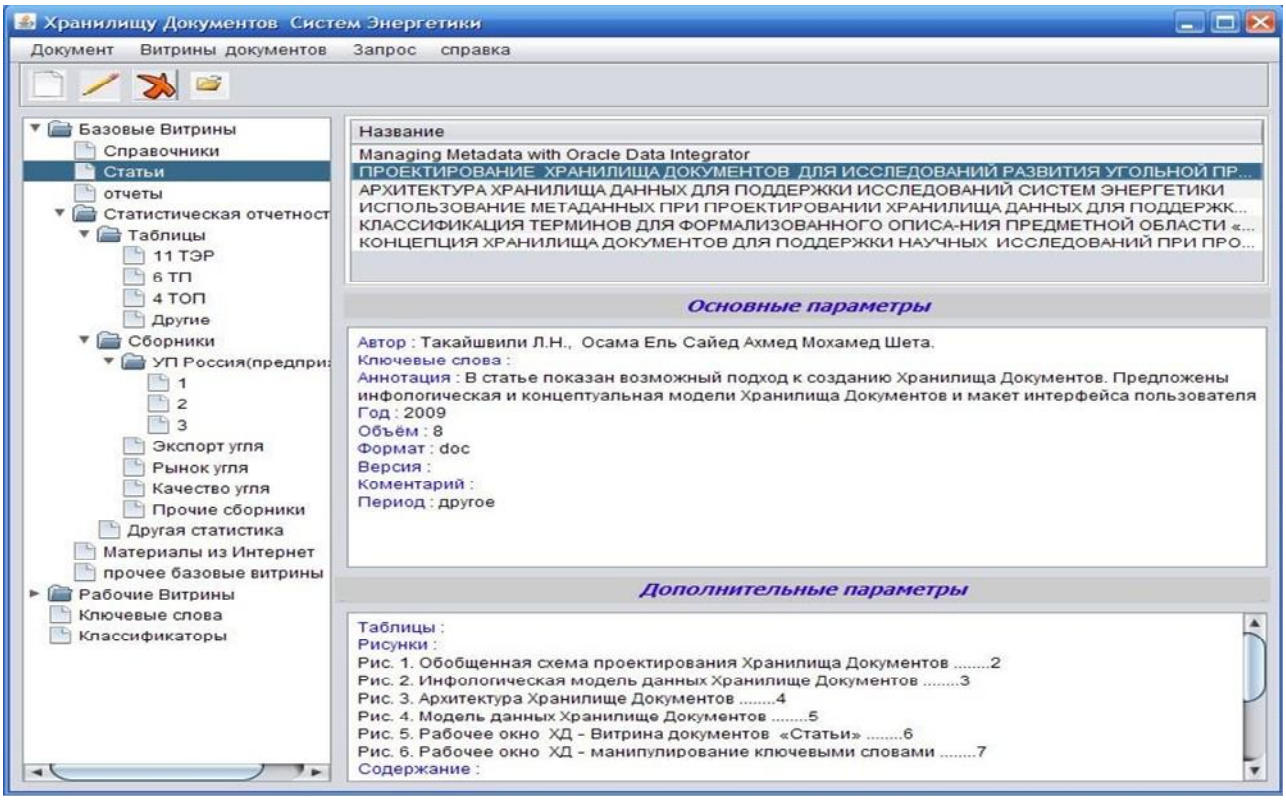


Рис. 8. Интерфейс программы администрирования хранилища знаний

В целом выполненная реализация хранилища знаний рассматривается как одна из возможностей интеграции информационных и интеллектуальных ресурсов для исследований энергетики.

Заключение. В статье описаны технологии использования хранилища данных и знаний для исследований отраслевых систем топливно-энергетического комплекса (ТЭК) и инструментальные средства поддержки этих технологий. Применение метаданных позволяет строить универсальные программные компоненты, взаимодействующие с хранилищем данных и знаний. Реализация инструментальных средств выполнена на объектно-ориентированном языке Java (Java Standard Edition) в среде NetBeans. Выбор языка Java обусловлен тем, что этот язык уже использовался при разработке информационной инфраструктуры, и реализация хранилища данных и знаний в Java-технологии облегчает его интеграцию в состав информационной инфраструктуры. В качестве базовой СУБД используется Firebird. Такой подход к созданию инструментальных средств и использованию метаданных дает возможность перейти практически на любую другую СУБД.

Разработка выполнена и используется в Институте систем энергетики им. Л.А. Мелентьева СО РАН. Работа частично поддержана грантами РФФИ №10-07-264, №08-07-00172 и грантом Программы Президиума РАН №2.29.

Автор выражает благодарность своему научному руководителю, д.т.н. Массель Л.В. и сотрудникам ИСЭМ СО РАН Копайгородскому А.Н. и Такайшвили Л.Н. за оказанную помощь в постановке задачи и выборе средств для ее решения, а также консультации в ходе выполнения работы.

Литература

1. Воропай Н.И., Массель Л.В. ИТ-инфраструктура системных исследований в энергетике и предоставление ИТ-услуг. – Известия АН – Энергетика, №3, 2006.- С. 86-93.
2. Копайгородский А.Н., Осама Ель Сайед Ахмед Мохамед Шета Архитектура хранилища данных для поддержки исследований систем энергетики // Труды XIV Байкальской Всероссийской конференции "Информационные и математические технологии в науке и управлении", том III. – Иркутск: ИСЭМ СО РАН, 2009. – С. 202-207.

3. Helen Borrie The Firebird Book: A Reference for Database Developers 2004 – 1092 с.
4. CMAP Tools – http://cmap.ihmc.us/
5. Маклаков С.В. Создание информационных систем с AIIFusion Modeling Suite. – М.: "Диалог-МИФИ", 2003. – 432 с.
6. Брюс Эккель. Философия Java (Thinking in Java).– 3-е изд.– СПб.: Питер, 2003. – 976 с.
7. Монахов В.В. Язык программирования Java и среда NetBeans.– 2-е изд.– СПб.: БХВ-Петербург, 2009.– 720 с.
8. Осама Ель Сайед Ахмед Мохамед Шета. Реализация хранилища документов для исследований энергетики // Труды XV Байкальской Всероссийской конференции "Информационные и математические технологии в науке и управлении", том III. – Иркутск: ИСЭМ СО РАН, 2010. – С. 103 - 110.