# О доступе к электронным коллекциям в виде реляционных баз данных на основе онтологий♣

© Е. В. Биряльцев, А. М. Гусенков, А. М. Елизаров

НИИ математики и механики им. Н. Г. Чеботарева
Казанского государственного университета

elizarov@ksu.ru

**Аннотация**

Работа посвящена применению онтологических описаний разных уровней – логической модели предметной области, модели представления данных и лексико-семантического тезауруса – для обеспечения доступа к электронным коллекциям в виде реляционных баз данных. Предлагаемые подходы апробированы на информационных ресурсах нефтегазовой индустрии.

## 1 Введение

Как известно, технологии управления информацией составляют в последние годы наиболее значимое и практически востребованное направление развития информационных технологий. Ключевую роль эти технологии играют и в электронных библиотеках, которые являются не только многочисленными апробированными и практически функционирующими информационными системами, реализующими какую-либо из технологий управления информацией (технологию баз данных (БД) или текстовых систем, веб-технологию), но и примерами использования разных сочетаний таких технологий в рамках одной системы. Как отмечено в обзоре [1], важной тенденцией последнего десятилетия в развитии технологий управления информацией стала их интеграция в конкретных реализациях информационных систем, а одним из развивающихся ее направлений – интеграция технологий баз данных с веб-технологиями. Технологии БД весьма важны для электронных библиотек, так как позволяют с помощью соответствующих СУБД создавать и обеспечивать эффективный доступ к разнообразным коллекциям структурированных данных (результатов наблюдений, измерений, научных экспериментов и т. д.), а создание пользовательских интерфейсов в системах бах данных на основе онтологий предметной области – признанная актуальная задача.

Существующие в настоящее время два наиболее популярных формализма представлений символьной информации – коллекции сплошных текстов и реляционные базы данных (РБД) – различаются по механизмам доступа к данным. Для реляционных баз данных запрос формируется в виде выражения на объектно-реляционном языке для систем баз данных SQL, для них характерны сложноорганизованный поиск и конструирование ответа на запрос из найденных элементов, раздельное индексирование по атрибутам. Для коллекций сплошных текстов запрос формируется в виде набора ключевых слов, реализуются поиск вхождения ключевых слов во фрагменты текстов (возможно, с учетом морфологии) и представление пользователю результатов в виде найденных документов. Для ускорения поиска применяется механизм индексирования текстов лексемами, содержащимися в них.

Простота задания запроса и универсальность механизмов поиска в сплошных текстах делают возможным применение критериев поиска, которые задаются пользователем, к произвольным коллекциям сплошных текстов. Доступ же к РБД осуществляется на основе предопределенных процедурных модулей с уникальной семантикой, что не позволяет сформулировать абстрактный запрос и применить его ко всем доступным базам данных. Это различие определило отсутствие в универсальных поисковых машинах механизмов поиска по реляционным базам данных, в результате чего затруднен доступ к огромным пластам информации. Поэтому пользователю, не знакомому со структурой и составом конкретных РБД, необходимо обеспечить возможность организации поиска на основе механизма ключевых слов. Ниже мы опишем, какие информационные структуры необходимы для решения этой задачи.

## 2 Особенности поиска по ключевым словам в РБД

Рассмотрим возможность организации поиска по ключевым словам в РБД только на основе информации самой базы данных. Современные реляционные СУБД хранят в своих системных

**Труды 9ой Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL'2007, Переславль-Залесский, Россия, 2007.**

данных наименования всех таблиц и столбцов. По этой информации и набору ключевых слов, заданному пользователем, достаточно просто автоматически построить набор SQL-запросов, выбирающих из таблиц базы кортежи, содержащие вхождения заданного набора ключевых слов в каждый из столбцов таблиц БД. Результатом поиска будет множество кортежей различного состава атрибутов, хотя бы в одном из которых встречается заданное ключевое слово. В итоге пользователь получит ответ в виде набора таблиц различного формата, столбцы которых в отсутствие других данных будут именоваться по названиям столбцов таблиц БД. С учетом специфики архитектуры и реализации РБД такое представление будет иметь следующие особенности:

- релевантность запросу записей каждой отдельно взятой таблицы будет невелика, в нормализованной РБД ключевые слова будут разбросаны по различным таблицам; в частности, практика формирования РБД такова, что все часто встречающиеся наименования организованы в отдельные таблицы-справочники;
- информативность наименований таблиц и столбцов для пользователя будет минимальной, названия столбцов в РБД фактически будут названиями программных объектов.

Естественно, что такая информация мало полезна конечному пользователю.

Отдельные таблицы в РБД, как правило, не обладают семантической законченностью; семантически законченные объекты (входные и выходные документы) конструируются из кортежей нескольких таблиц, связанных между собой общими ключами. Релевантность ответа может быть существенно повышена при поиске вхождений ключевых слов не в произвольные таблицы БД, без учета их связей между собой, а в наборы документов.

Неинформативность наименований артефактов баз данных требует хранения развернутых наименований, ориентированных на восприятие их человеком. Наличие таких наименований, ориентированных на пользователя, позволяет расширить область поиска ключевых слов также на наименования, что очень полезно, так как граница между наименованием столбца и значением столбца в базах данных подвижна, и в одной схеме искомая лексема может находиться среди значений, а в другом – среди наименований (см. [2]).

Следующие две взаимосвязанные проблемы, возникающие при поиске по ключевым словам в базе данных, возникают и при поиске по ключевым словам в сплошных тестах. Для реляционных баз данных они приобретают большую остроту. Первая проблема связана с возможными различиями в уровне абстракции представления информации. Например, в одной из баз данных искомая информация может содержаться в таблице «Детали машин», в другой – в отдельных таблицах «Колеса», «Насосы» и т. д. Различаться может и уровень абстракции самого набора ключевых слов, задаваемых различными пользователями как поисковый критерий. Вторая проблема связана с возможными различиями терминологии, используемой разработчиками базы данных и пользователями. Помимо обычной синонимии терминологические различия могут выражаться в употреблении различных сокращений и аббревиатур, которые, как показывает анализ реальных текстов в базах данных [2], используются в БД чаще, чем в сплошных текстах.

Острота этих проблем для РБД обусловлена характером хранения информации в них. Правила нормализации и практика использования искусственных ключей приводят к тому, что значение атрибута хранится в явном виде только один раз, а при необходимости его повторного употребления используется ссылка. В сплошных текстах атрибут, о котором идет речь, именуется преимущественно явно, а использование местоимений (некоторый аналог ссылки в РБД) ограничено. При неоднократном явном именовании атрибута в сплошном тексте более вероятны применение синонимов, упоминание гипонимов и гиперонимов для пояснения оттенком смысла и т. п., расширение лексико-семанти-ческого поля. В хорошо спроектированных РБД лексико-семантическое поле значительно беднее.

Указанные факторы резко снижают полноту, информативность и релевантность выборок, полученных при поиске вхождения ключевых слов в таблицы баз данных. Для повышения названных характеристик до минимально необходимого уровня поисковые машины должны иметь доступ к вспомогательным информационным ресурсам и строить поиск с их использованием.

## 3 Вспомогательные информационные ресурсы

Как было сказано выше, для эффективного поиска по ключевым словам в реляционной базе данных поисковому механизму не хватает информации нескольких типов.

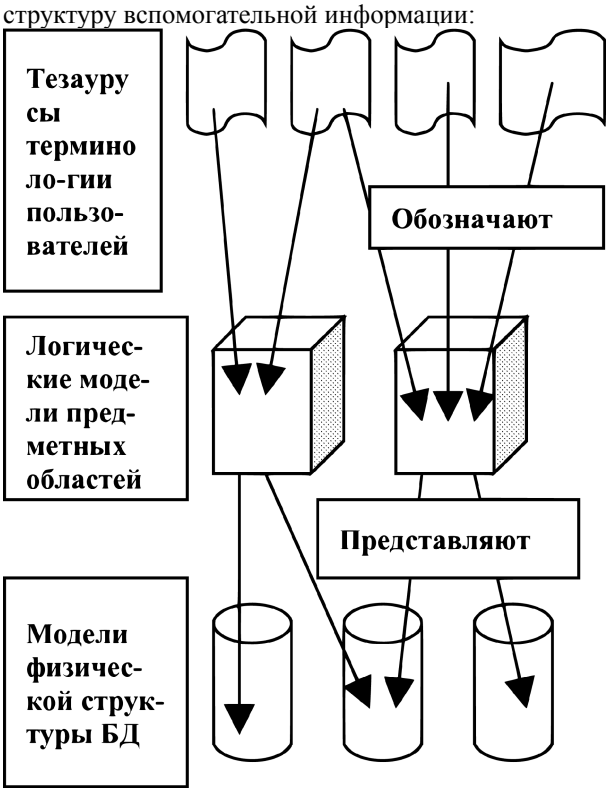
Во-первых, это информация о структуре самой базы данных. Связность и близость значений атрибутов БД определяются связностью кортежей через общие ключи. Таким образом, для выдачи пользователю осмысленной информации поисковая машина должна иметь информацию о возможных соединениях кортежей по ключам. Такую информацию достаточно легко получить непосредственно из схем без данных, так как современные СУБД хранят информацию о ключах в своих системных данных. Восстановление всех возможных осмысленных ком-бинаций таблиц базы (документов) также является достаточно простой задачей, как показано ниже.

Во-вторых, это семантические маркеры для единообразной разметки атрибутов всех баз данных. Для маркирования удобно использовать логическую модель предметной области со связями

«класс – подкласс – экземпляр» и «часть – целое». Такой подход известен как Semantic Web [3] и в настоящее время широко апробируется в различных приложениях, в частности, для повышения релевантности поиска в сплошных текстах. РБД фактически уже размечены наименованиями атрибутов и в этом отношении имеют большую степень готовности к использованию совместно с технологиями Semantic Web, так как необходимо установить соответствие между наименованием атрибута и некоторой семантической единицей.

В-третьих, поисковый механизм должен содержать лексико-семантическую информацию о терминологии, применяемой различными группами пользователей для обозначения тех или иных объектов, свойств и связей в предметной области (пользовательские тезаурусы). Наличие такой информации позволит решить проблемы синонимии, сокращений и аббревиатур, применяемых различными группами пользователей.

Таким образом, мы имеем следующую структуру вспомогательной информации:




Следует отметить, что каждая из этих информационных структур может быть представлена во множестве экземпляров. Поисковый механизм может иметь доступ ко множеству баз данных, эти базы могут быть связаны с различными предметными областями или быть междисцплинарными, различные группы пользователей могут использовать различные терминологические тезаурусы, вплоть до мультиязычных.

Возникает вопрос о формализме представления столь разнородной и объемной информации. Каждая из этих моделей имеет некоторый устоявшийся формализм представления. Так, схемы баз данных традиционно представляются с использованием SQL DDL, логические модели – ER-диаграммами, лексико-семантическая информация – тезаурусами. Очевидно, что оперирование столь разнородными моделями технически затруднительно. Для наших целей оптимально использовать единый достаточно мощный формализм, ориентированный на автоматическую обработку информации, хранящейся в моделях. В качестве такового можно предложить формализм онтологий [4, 5]. Онтологии включают как частный случай логические и физические модели данных и тезаурусы, так что с теоретической точки зрения мощности формализма онтологий достаточно для представления всех трех типов информационных моделей.

При использовании совокупности онтологий, отражающих различные аспекты представления знаний, удобно применять и единый язык их описания. В качестве такового можно использовать язык OWL, разработанный рабочей группой Semantic Web Activity и рекомендованный консорциумом W3C, а именно, диалект языка OWL-DL (Description Logic) [6]. Выбор языка OWL-DL обусловлен чисто практическими аспектами, в частности, его поддержкой в существующих сегодня системах описания знаний и системах логического программирования, а также перспективами его распространения в будущем как международного стандарта.

Таким образом, вспомогательные информационные ресурсы содержат физические модели баз данных, логические модели предметной области и тезаурусы пользовательской терминологии, представленные в формализме онтологий.

Общая схема использования этих структур поисковым механизмом следующая. Наборы ключевых слов, задаваемые пользователем, приводятся к семантическим объектам логических моделей предметной области, по которым уже производится поиск в конкретных РБД с использованием маркировочных связей. Естественно, такая схема требует предварительной настройки связей, что, в принципе, не более трудоемко, чем семантическое маркирование сплошных текстов. Мы не будем рассматривать вопросы реализации процесса установления связей, а более подробно обсудим, что представляют собой данные онтологии и каким образом возможно их построение.

## 4 Онтология баз данных

Известны подходы к описанию структуры конкретных реляционных баз данных с помощью онтологий для конкретных предметных областей (например, [7]). Применительно к решаемой нами задаче можно предложить следующее универсальное представление структуры реляционных БД в формализме онтологий. Основой являются универсальные (не зависящие от

конкретной базы данных) концепты ТАБЛИЦА, СТОЛБЕЦ, КЛЮЧ, ДОМЕН, соответствующие основным объектам баз данных, и универсальные отношения между ними:

ТАБЛИЦА *содержит* СТОЛБЕЦ
ТАБЛИЦА *имеет первичный* КЛЮЧ
ТАБЛИЦА *имеет внешний* КЛЮЧ
КЛЮЧ *содержит* СТОЛБЕЦ
СТОЛБЕЦ *имеет тип* ДОМЕН (1)

Объекты (таблицы, столбцы, ключи и домены) конкретной базы данных во втором случае представляются как экземпляры универсальных концептов соответствующего типа.

Любая информация в реляционной базе данных имеет вид набора таблиц, возможно связанных между собой общими ключами. Для анализа этих связей необходимо последовательное нахождение способов извлечения из базы данных ее атрибутов в виде таблиц, задаваемых пользователем. В формализованном виде поставленная задача имеет следующий вид (назовем ее задачей А): *даны онтология О вида (1) и произвольное множество {С} столбцов из О; требуется определить, возможно ли построение {С}, и если возможно, то каким образом.*

Если в одной из таблиц РБД требуемое множество столбцов содержится как подмножество, то задача решается тривиально. Однако в реальных случаях искомое множество {С} будет содержаться, вероятнее всего, в нескольких различных таблицах. Для восстановления искомого множества из реально существующих таблиц требуется соединение этих таблиц между собой. Правила манипулирования реляционными данными разрешают соединения таблиц только при наличии общего ключа, причем в одной из таблиц этот ключ играет роль первичного ключа, а в другой – вторичного. Для формализации этих правил введем в онтологию (1) две функции интерпретации (ФИ) следующего вида:

ФИ1: Если ТАБЛИЦА1 имеет первичный КЛЮЧ1 и ТАБЛИЦА2 имеет внешний КЛЮЧ1, то существует ТАБЛИЦА3, содержащая столбцы, принадлежащие ТАБЛИЦА1 и ТАБЛИЦА2;

ФИ2: Если ТАБЛИЦА1 содержит СТОЛБЕЦ1, то существует ТАБЛИЦА2, содержащая все остальные столбцы ТАБЛИЦА1, кроме СТОЛБЕЦ1.

Первая функция интерпретации соответствует операции соединения по ключу, вторая – операции проекции реляционного отношения, необходимой для сокращения получаемого при соединении таблиц множества столбцов до искомого.

Задача А сводится к нахождению такой последовательности применения ФИ1 и ФИ2, которая даст в качестве результата искомое множество {С}. Задача сводится к известному классу задач о блуждании по ориентированному графу, где вершинами являются экземпляры концепта ТАБЛИЦА, а дугами – наличия общего ключа, ориентированные посредством отношения «*имеет первичный*» и «*имеет вторичный*». Подходы к решению такого класса задач хорошо известны [8] и представляют лишь вычислительную сложность при большой размерности задачи.

## 5 Построение онтологии предметной области на основе логических моделей данных

Основной в паре «онтология предметной области – онтология профессионального языка» является онтология предметной области. Несмотря на активное создание в последнее время онтологий разнообразных предметных областей, они, как правило, носят экспериментальный характер. Онтологии, которые являлись бы для какой-либо предметной области стандартом де-юре или де-факто, в настоящее время отсутствуют. В качестве прототипа онтологии предметной области можно использовать логические модели, созданные в некоторых отраслях и имеющие статус отраслевого стандарта. Одной из них является модель данных Epicentre версии 3.0 Нефтетехнической корпорации Petrotechnical Open Software Corporation (POSC) [9]. Она представлена в виде ER-диаграмм [10], а также в виде набора текстовых файлов на языке EXPRESS (ISO 10303, part 11). Это представление ориентировано на генерацию структур баз данных по логической модели, а также на визуальное восприятие ИТ-специалистами. Априори возможность представления названной модели в виде онтологий была неочевидной.

В модели данных Epicentre определено более 1000 реально существующих технических и бизнес-объектов, связанных с разведкой и добычей нефти. В терминологии POSC-моделирования данных эти объекты названы сущностями (entities). В модели определены характеристики, которые могут содержать сущности, названные атрибутами сущностей (attributes). Наиболее важными являются атрибуты, определяющие взаимосвязи между сущностями.

Один из основных принципов модели Epicentre основан на различии между объектами, свойствами или характеристиками объектов (properties) и видами деятельности (activities), которые используют объекты и определяют их свойства. Отличительная особенность свойств объекта – это возможность иметь многократные версии или описания, а также однозначная связь свойства со своим собственным определением (описанием) или историей обработки. Модель Epicentre рассматривает принцип моделирования, при котором конкретному экземпляру сущности – объекту – обеспечивается его существование. Поэтому каждый экземпляр представляет отдельный объект и не может быть версией существующего объекта. Версии характеристик объекта отделены от объекта. В Epicentre эти характеристики связаны с наличием атрибутов или свойств сущностей. Также в модели Epicentre спецификация сущностей была расширена для того, чтобы включить сущности, имеющие стандартный набор экземпляров. Подобные сущности называются справочными (reference_entities) и отличаются от других тем, что

некоторый стандартный набор значений был определен POSC. Их присутствие требуется для совместимости со спецификациями POSC. Все справочные сущности имеют дополнительные характеристики, позволяющие задать источник информации, содержащейся в экземпляре, и связанную с ним библиографию.

Другая фундаментальная часть архитектуры Epicentre – это понимание того, что многие сущности характеризуются пространственным представлением. Чтобы облегчить общее использование пространственных данных для всей модели, POSC определил набор общих пространственных объектов и пространственных связей. Для каждого геометрического объекта деятельности в разных разделах модели может быть задано его местоположение через отношения с одним или несколькими общими пространственными объектами.

При построении OWL-онтологии были использованы следующие основные подходы:

- любой сущности Epicentre соответствует простой именованный класс OWL-онтологии с сохранением в именах классов приставок, позволяющих идентифицировать сущности-свойства и сущности-справочники; все эти классы располагаются в корне таксономического дерева онтологии;
- степени связности сущностей (один-к-одному, один-ко-многим, многие-ко-многим) в OWL соответствует определение простых свойств-атрибутов, если связанная сущность не является типом данных, и свойств-значений в противном случае; указание степени связи между классами реализовано с помощью понятия кардинальности в OWL.

В OWL отсутствуют структурные элементы, которые в полном объеме описывают определение уникальности Epicentre. Поэтому в определение каждого класса на языке OWL добавлено новое предопределенное свойство, в котором перечислены все атрибуты, образующие уникальный ключ. Аналогичным же образом решена проблема сохранения условий ограничений.

Для каждой категории данных Epicentre в OWL-онтологии построены отдельные классы, в свойствах которых использовались встроенные типы данных языка OWL. Построена формальная LR(1) грамматика модели Epicentre, на основе которой реализовано семантическое преобразование модели Epicentre версии 3.0, описанной на языке EXPRESS, в онтологию на языке OWL. Выполнена русификация описания сущностей и атрибутов модели Epicentre, а также соответствующих им классов и свойств на OWL. Более подробно схема построения OWL-онтологии на основе модели Epicentre изложена в [11]. Онтология реализована на диалекте языка OWL DL, соответствующем правилам дескриптивной логики, что в дальнейшем позволит использовать системы логического вывода на экземплярах понятий.

Как сказано выше, второй необходимой компонентой является онтология профессионального языка, в терминах которого определяются наименования артефактов баз данных и онтологии предметной области. Анализ реальных баз данных показывает распространенность использования в наименованиях аббревиатур, сокращений, лексических конструкций бытовой лексики в переносном смысле и т. п. Все эти особенности делают необходимым создание отдельной лингвистической онтологии, описывающей лексико-семантические отношения.

## 6 Лингвистическая онтология

Для создания лингвистической онтологии природно-технических объектов выбран подход, основанный на построении тезаурусов WordNet [12, 13]. Словарь предметной области был построен путем объединения словоформ из описаний сущностей и атрибутов модели Epicentre и описаний атрибутов таблиц и доменов таблиц-справочников реляционных баз данных нефтедобывающей корпорации ОАО «Татнефть». Лексико-семантические характеристики баз данных, использованных при построении словаря, описаны в [2, 14]. В настоящее время словарь содержит около 6000 словоформ. Для каждого слова определяется входной синонимический ряд (синсет). На лексико-семантических вариантах слов и синсетах определены следующие отношения: гипонимия, часть – целое, несовместимость, антонимия, конверсивность, омонимия.

Описаниям синсетов построенного таким образом тезауруса присущи особенности, вытекающие из специфики описания атрибутов реальных баз данных: наличие коротких фраз, сокращений, технических аббревиатур, орфографических ошибок (см. [2]). Однако в отличие от анализа сплошных текстов для баз данных имеется возможность уточнения распознавания входных слов путем просмотра содержимого доменов атрибутов таблиц и сопоставления их с онтологией предметной области.

## Заключение

В работе обсуждены подходы, обеспечивающие доступ к данным в реляционных базах данных на семантическом уровне. В частности, рассмотрены возможности поиска с учетом декомпозированного представления текстовых документов с распределением их по столбцам различных таблиц базы данных, с использованием семантической разметки документов и онтологий предметной области. Описано применение предлагаемых подходов для предметной области разведки нефтяных месторождений и добычи нефти.

Представленные выше результаты получены в рамках проекта РФФИ 06-07-89219,

основной целью которого является разработка подходов к автоматизации построения интеграционных процедур. Вместе с тем, обеспечение произвольного поиска в реляционных базах данных имеет более широкий круг применений, чем интеграция РБД, что позволяет рассматривать его как важную самостоятельную задачу.

## Литература


[1] Когаловский М. Р. Тенденции развития технологий управления информационными ресурсами в электронных библиотеках// Тр. 8ой Всерос. научн. конф. «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2006, Суздаль, 17 – 19 октября 2006 г. – Ярославль: Ярославский гос. ун-т им. П. Г. Демидова, 2006. – С. 46-55.

[2] Биряльцев Е. В., Гусенков А. М., Галимов М. Р. Особенности лексико-семантической структуры наименований артефактов реляционных баз данных // Тр. Казан. школы по компьютерной и когнитивной лингвистике TEL'2005. – Казань: Изд-во Казан. ун-та, 2006. – Вып. 9. – С. 4-12.

[3] Semantic Web. – http://www.w3.org/2001/sw/.

[4] Гаврилова Т. А., Хорошевский В. Ф. Базы знаний интеллектуальных систем. – СПб.: Питер, 2000. – 384 с.

[5] Добров Б. В., Иванов В. В., Лукашевич Н. В., Соловьев В. Д. Онтологии и тезаурусы: Учебно-методическое пособие. – Казань: Изд-во Казан. ун-та, 2006. – 198 с.

[6] W3C World Wide Web consortium. – http://www.w3.org/TR/2004/REC-owl-guide-20040210/.

[7] Жучков А.В. и др. Новые технологии для понятийных сетей, создаваемых в рамках МНТП «Вакцины нового поколения и диагностические системы будущего» // Электронные библиотеки. – 2003. – Т. 6, вып. 6. – http://www.elbib.ru/index.phtml?page=elbib/rus/journal/2003/part6/ZATGS.

[8] Джеймс А. Андерсен. Дискретная математика и комбинаторика: Пер. с англ. // М.: Издательский дом «Вильямс», 2003 – 960 с.

[9] Petrotechnical open standards consortium. – http://www.energistics.org/posc/General_Stds.asp?SnID =188137941.

[10] Дейт К. Д. Введение в системы баз данных. 7-е изд. – М.: Вильямс, 2001. – 702 с.

[11] Биряльцев Е. В., Гусенков А. М., Хайруллина А. И. Представление модели данных Epicenter POSC на языке онтологий OWL // Тр. Казан. школы по компьютерной и когнитивной лингвистике TEL-2006, Казань, 2007.

[12] Fellbaum C. (ed.) WordNet: An Electronic Lexical Database. – MIT Press, 1998.

[13] Vossen P. Building a multilingual database with wordnets for several European languages. – http://www.illc.uva.nl/EuroWordNet/.

[14] Биряльцев Е. В., Гусенков А. М., Косинов Я. Г. Представление структуры реляционных баз данных в формализме онтологий // Тр. Казан. школы по компьютерной и когнитивной лингвистике TEL-2006, Казань, 2007.


## About access to electronic collections presented as relational databases on the basis of ontologies


E. V. Birialtcev, A. M. Gusenkov, A. M. Elizarov



This paper is devoted to application of ontological descriptions of different levels for providing access to electronic collections presented as relational databases. Offered approaches are approved on information resources of the Oil&Gas industry.



♣Работа выполнена при поддержке РФФИ (проект № 06-07-89219) и РГНФ (проект № 07-01-12146)