СИСТЕМАТИЗАЦИЯ ФИЗИКО-ХИМИЧЕСКИХ ДАННЫХ. ВОЗМОЖНОСТИ ОНТОЛОГИЙ И БАЗ ДАННЫХ, ПРИНЦИП ДОПОЛНИТЕЛЬНОСТИ

В.А. Серебряков[1], А.О. Еркимбаев[2], В.Ю. Зицерман[2], Г.А. Кобзев[2], К.Б. Теймуразов[1], Р.И. Хайрулин[2]

[1] Вычислительный центр им. А.А.Дородницына РАН
[2] ОИВТ РАН

Накопление и систематизация физико-химических данных и прежде всего данных о свойствах вещества является одним из основных направлений деятельности в естественнонаучных дисциплинах. До недавнего времени основным информационным ресурсом этой области была традиционная БД. Однако резкий рост их числа при многообразии форматов и моделей привел к тому, что действующая инфраструктура оказалась неспособной обеспечить эффективную организацию рабочего процесса. Автономность БД с жесткой фиксацией используемых терминов и логических структур практически исключает возможность беспрепятственного обмена данными без активного участия человека-эксперта. В итоге интерес в научном сообществе сместился в сторону принятия онтологий как более эффективных средств формализации и распространения научных данных. Главное назначение онтологий в том, что они добавляют к данным семантику (смысл, правильное толкование) и соотношения, что в совокупности описывает «знание» с возможностью его машинной интерпретации. Оценить богатый потенциал, заложенный в онтологиях для хранения и распространения данных, можно, сопоставляя их возможности с БД [1-3]. При кажущемся сходстве решаемых задач между ними имеются глубокие различия. Онтология служит для распространения информации, определяя на формальном языке концепции и соотношения, которые представляют содержание и структуру предметной области. В то же время концептуальная схема БД, определяя все понятия и структуру данных, служит только для тех целей, что реализует конкретная БД. Распространение информации посредством онтологии проводится согласованным образом, то есть передаваемая ею структура данных является общедоступной и одинаково трактуемой в определенном сообществе. Все члены сообщества могут использовать онтологию и имеют доступ к информации.

Авторы ряда публикаций специально выделили все факторы, определяющие сходства и различия обеих конструкций. Их сводка, зафиксированная в лекции [2], приведена в таблице 1.

Таблица 1.

Концептуальная схема БД	Онтология
Определяет структуру БД на формальном языке	Определяет набор концепций и соотношений, которые представляют содержание и структуру предметной области на формальном языке
Фокусируется на данных	Фокусируется на смысле
Сущности	Классы
Атрибуты	Соотношения
Ограничения	Аксиомы
Нет таксономии	Таксономия – ключевой элемент онтологии
Данные – ключевой элемент	Экземпляры данных не обязательны
Семантика только в концептуальной схеме, спроектирована для человека, не эволюционирует с изменением БД и приложений	Семантика – ключевой элемент, доступный программной обработке
Схему трудно изменять и поддерживать	Потенциально легче изменять и поддерживать

Наряду с семантикой, которая теряется в ходе проектирования БД, в таблице указаны некоторые дополнительные признаки. В частности, БД немыслима без содержания в виде данных, в то время, как в онтологии наличие экземпляров не обязательно (*instances optional*). С другой стороны, онтология немыслима без таксономии классов (хотя и не сводится к ней), а в БД таксономия как структурный компонент отсутствует. В итоге, формальный характер онтологии позволяет реализовать машинные выводы и рассуждения, что совершенно не предусмотрено в концептуальной схеме БД. Применительно к тематике «свойства веществ и материалов» особую роль играет возможность на уровне онтологий поддерживать эволюцию схемы данных, связанную с расширением круга объектов и появлением новых, ранее неизвестных понятий (последний пункт в табл.). Так, эволюция схемы может быть связана с постепенным включением наноструктур в химические БД [4, 5].

Поскольку БД отличает высокая производительность при поиске и реализации сложных запросов, недостижимая для других архитектур, в качестве основной стратегии выбрано не «вытеснение» БД, а создание своеобразной связки путем так называемого *database-to-ontology mapping.* Ее задача — использовать преимущества обеих конструкций, в основном за счет соединения семантики с высокой производительностью при работе с данными. В известном смысле, выбранный путь следует «принципу дополнительности»

при использовании обеих концепций, обладающих в чем-то взаимоисключающими свойствами.

Авторы [3], исходя из того, какая из двух концепций (онтология или схема БД) может рассматриваться как ведущая, выделили две стратегии. Первая из них ориентирована на использование онтологий для усиления функциональности БД: реализации запросов с использованием согласованной в сообществе специалистов семантики; облегчения проектирования БД; обеспечения интеграции нескольких БД. Вторая стратегия, напротив, рассматривает БД как возможный инструмент усиления или даже разработки онтологии.

Соответственно можно выделить две концепции взаимного отображения БД и онтологии. Первой соответствует БД, основанная на онтологии (**DBBO**, *database based on ontology*), когда смысл сущностей, записанных в БД, определен ссылкой на соответствующую онтологию. Вторая концепция предполагает проектирование структуры, получившей название онтологии, основанной на БД (**OBDB**, *Ontologies Based on DB*). Задача такой структуры — разместить экземпляры онтологии в БД, чтобы обеспечить достаточную эффективность при загрузке, поиске и реализации сложных запросов. В основном эта концепция нацелена на решение задач Semantic Web, где требуется организовать хранение и управление архивами документов, записанных на языке **RDF,** в то время как первая концепция (**DBBO**) обращена, в основном, к сообществу специалистов по БД.

Ниже речь будет идти о системах типа **DBBO**, включающих: исходно принятую БД с заполняющими ее записями; онтологию для семантического индексирования БД; возможные ссылки на другие онтологии с целью расширения словаря; соотношения между каждым элементом БД и онтологическим понятием. Онтология в такой структуре лишь обеспечивает семантику, но не содержит экземпляров — их роль выполняют записи БД.

Усиление функциональности за счет онтологии возможно не только для существующей БД, но и на этапе проектирования, что заметно ускоряет разработку и снимает ряд возникающих ограничений. В частности, концептуальные схемы, предлагаемые разными экспертами, в известной степени унифицируются, поскольку базируются на единой модели области. Кроме того, уже на этапе проектирования снимается проблема семантики, поскольку каждый элемент данных получается отображением онтологических понятий, имеющих точное определение и смысл.

Наряду с проектированием БД, использование онтологий открывает возможности в интеграции БД, ранее созданных без ссылок на онтологии. Онтология предоставляет достаточно эффективный способ объединения схем, позволяя преодолеть проблемы с отсутствием семантики в исходных БД. Пользователь может формулировать в терминах онтологии запросы, которые будут обращены к «подключенным» к онтологии БД. Тем самым онтология будет играть роль эффективного посредника между пользователем и данными.

Один из наиболее впечатляющих примеров расширения функциональности БД за счет связывания с онтологией дает химическая БД низкомолекулярных соединений содержания **ChEBI** (Chemical Entities of Biological Interest Ontology) [5], которая охватывает молекулярные сущности, их группы и классы. Первое из понятий относится к любым идентифицируемым по составу атомам, молекулам, наноструктурам и т.п., второе — группу связанных атомов (или один атом) в составе молекулы (methyl, CH3), третье — (класс) совокупность молекул или групп, охваченных классификационным признаком, например alkane (R-CH3). Каждую сущность, а также группу и класс , идентифицирует уникальный идентификатор **ChEBI** ID (например, **CHEBI:15377** для воды), свободный для цитирования в сети пользователем или программным агентом.

В записи для молекулярной сущности (рис.1) указаны химические данные (определение, формула, масса, заряд), названия, как принятое в **ChEBI**, так и синонимы из других источников, структурная информация, кодированная в линейных нотациях (**InChI** и **SMILES**), а также детализированная в **Molfile** (координаты атомов и матрица связности). Наряду с этим в записи представлены регистрационные номера в химических классификаторах (CAS, Beilstein и др.) и ссылки на другие БД с обширной информацией, например термодинамической в БД NIST (webbook.nist.gov/chemistry/) или токсикологической в БД ChemID*plus* (chem.sis.nlm.nih.gov/chemidplus/chemidlite.jsp).

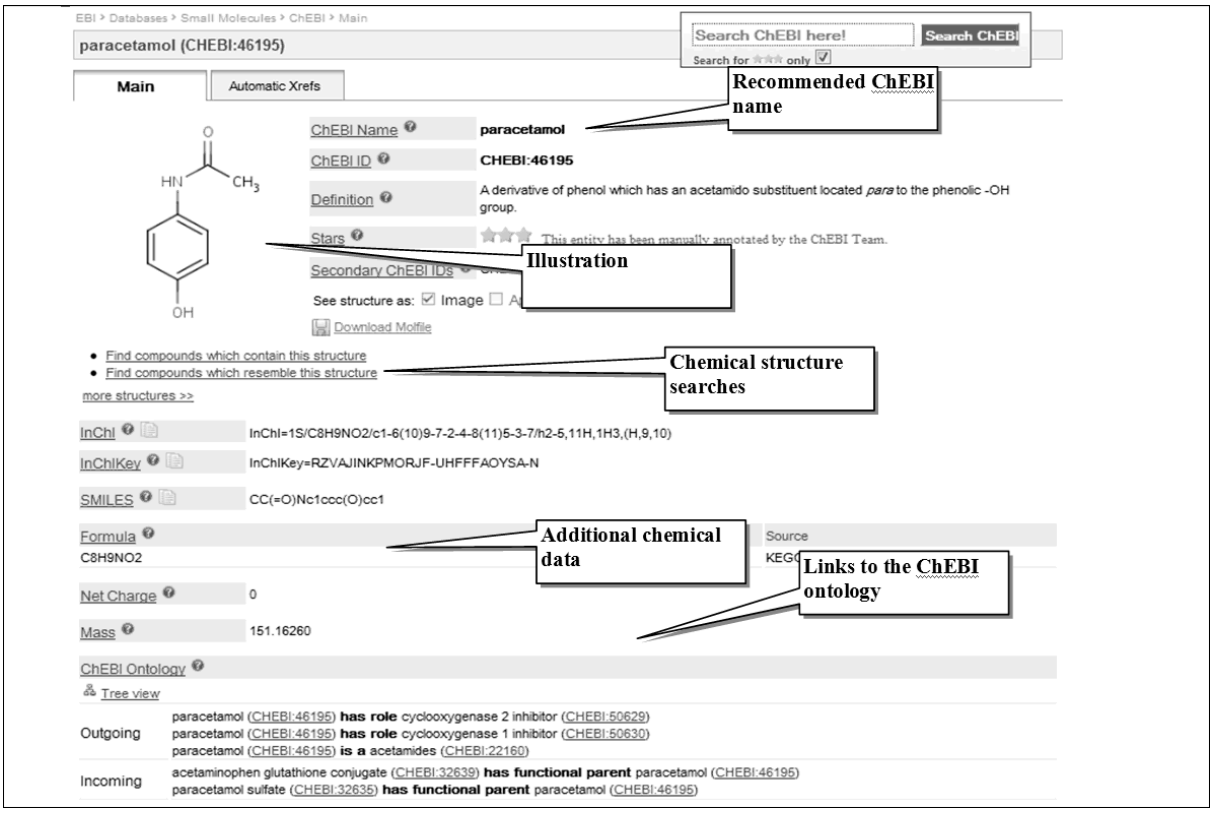

Рис. 1. Внешний вид типовой записи в БД **ChEBI.**

Главным элементом онтологии, интегрированной с БД, является таксономия, причем любое понятие может происходить от нескольких родительских понятий. Внизу рис. 2 показаны дочерние (incoming) и родительские (outgoing) классы, непосредственно связанные с веществом в записи (в данном случае **paracetamol, C8H9NO2**). Обращение к URI с указанием онтологии и уникального ID непосредственно связывает внешний ресурс (или пользователя) с соответствующей записью, например URI www.ebi.ac.uk/chebi/searchId.do?chebiId=46195 связывает его с записью в БД для парацетамола.

Отдельные классы связаны соотношениями, либо типа класс-субкласс (таксономия), либо ассоциативными соотношениями, которые определяют свойства или роли отдельных понятий, см. табл. 2.

Таблица 2.

Таксономические соотношения	**Is_a**
	Has_part
Ассоциативные соотношения	**Is_conjugate_base_of**
	Is_conjugate_acid_of
	Is_tautomer_of **Is_enantiomer_of**
	Has_functional_parent
	Has_parent_hydride
	Is_substituent_group_of
	Has_role

Первые два (**is_a, has_part**) определяют простейшие виды логической связи, восемь остальных — фиксируют химический контекст в отношениях двух веществ, например, соотношение **has_functional_parent (**как видно из рис. 2) указывает, чем является ***paracetamol*** по отношению к другому веществу ***paracetamol sulfate***. Особое значение в списке имеет последнее соотношение **has_role**, позволяя раскрыть множество аспектов в свойствах и приложении веществ. Оно связывает базовую суб-онтологию **Molecular structure**, в рамках которой построена таксономия веществ, с другой суб-онтологией **Role**, которая включает три класса верхнего уровня **(biological role, chemical role, application**). При этом **chemical role** классифицирует сущности по их «химически значимой» роли (кислота, основание, лиганд), **biological role** определяет их «биологически значимую» роль, а **application** дает классификацию по целевому использованию, например **drug, pesticide** и др. Все сущности, относящиеся к этой суб-онтологии, так же как и вещества, имеют уникальный ID, что позволяет связывать внешний ресурс с любой из сущностей в онтологии **role**. Пользователь может совершать навигацию по

дереву, соответствующему ролевой онтологии, точно так же, как и для веществ, включенных в БД.

Таким образом, наличие онтологии открывает для пользователя или внешнего ресурса возможности, исходно отсутствующие в БД. В частности, по идентификатору **ChEBI** можно сделать ссылки на любые сущности БД, определить их логические и ролевые связи, встроенные в многоуровневые таксономии. В результате можно возложить на программного агента реализацию сложных запросов с поиском веществ по структуре, функциям, биологической роли и т.п. В целом, это означает, что наряду с данными для конкретного вещества, **ChEBI** предоставляет фрагмент «знаний», характеризующих предметную область.

Общая концепция **DBBO** была опробована ранее при создании БД по теплофизическим свойствам веществ [6]. Роль онтологии здесь особенно существенна, поскольку наряду с унификацией семантики необходимо поддерживать эволюцию схемы данных и аксиомы, отражающие логические и математические ограничения, присущие данной области. Ключевое понятие — набор данных, включающий для одного вещества несколько констант и температурных функций, а также сведения о фазовом состоянии вещества, единицах измерений, неопределенности и источнике данных. Основные списки — веществ, свойств, фазовых состояний, единиц измерений и т.д. считаются открытыми, что позволяет в рамках онтологии поддерживать эволюцию схемы данных.

Концептуализация предметной области привела к выбору 12 базовых понятий, послуживших основой для построения соответствующих классов. Среди них группа базовых классов (вещества, состояния, свойства, численные данные), 6 вспомогательных классов (размерности, неопределенность, источник данных и др.) и 2 класса, определяющих вычисляемые функции и аргументы. При этом класс **Functions** порождает 2 суб-класса, определяющих значения свойств и выполнение математических ограничений, определяемых требованиями предметной области, например равенство энергии Гиббса для сосуществующих фаз (жидкость-газ или жидкость-твердое тело).

При разработке онтологии активно привлекались внешние источники (существующие онтологии и словари) для унификации семантики. В частности, для именования веществ использован словарь **ChemSpider**, который обеспечивает присвоение веществу уникального идентификатора, например для водорода **CSID:762** и соответствующего URI www.chemspider.com/Chemical-Structure.762.html. Отдельные термины, связанные с фазами, свойствами, размерностями приняты из онтологий **ChemAxiom** и **QUDT** (*Quantities, Units, Dimensions and Data Types*) [4].

Физические принципы, определяющие свойства веществ, накладывают целую совокупность ограничений на использование понятий. Логические ограничения записаны с использованием конструкций языка OWL. Среди логических ограничений — разбиение класса свойств на два непересекающихся

класса (свойства-функции и свойства-константы); обязательность определения аргумента для свойства-функции; согласованность ссылок на состояния вещества с видом свойств-функций (например, запрет на свойство **viscosity** в состоянии **solid**). Математические ограничения задаются отдельно для каждого экземпляра-свойства. Они относятся к свойствам-функциям: требования к области определения, области существования, характеру монотонности и возможно другим характеристикам (например, связи двух и более функций). Математические ограничения касаются не классов, а экземпляров свойств, поскольку области определения и существования функций определяются для каждого свойства отдельно и список свойств допускает расширение. На основании построенной онтологии сгенерирована реляционная БД в СУБД PostgreSQL. Соответственно 12 классам онтологии создано 12 Java классов, которые отображаются на таблицы реляционной БД данных, рис. 2.

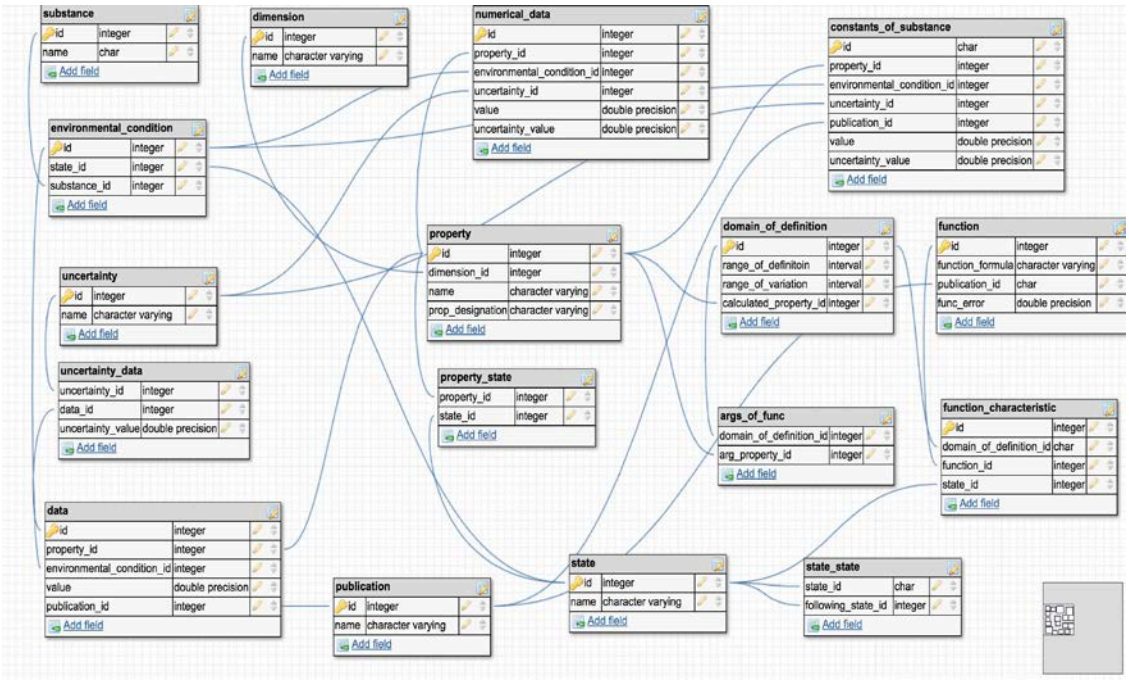

Рис. 2. Таблицы теплофизической БД [6].

Разработанная система заметно облегчила проектирование реляционной БД за счет связи с онтологией. Помимо унификации семантики, онтология обеспечила выполнение логических и математических связей между понятиями и сохранение за пользователем права наращивать списки веществ, свойств, единиц измерения и прочих элементов набора данных при эволюции концептуальной схемы.

В работе изучен один из новых подходов к работе с данными по свойствам, использующий технологии связывания БД с онтологиями. Сами по себе БД, будучи созданы в разных коллективах, с неизбежностью порождают разнобой в терминологии, логических схемах и форматах данных. Онтология оказалась идеальным средством унификации семантики и подчинения

множества концептуальных схем единой структуре, к тому же согласованной в научном сообществе. В результате открывается путь к широкой интеграции структурно или даже тематически разнородных ресурсов с возможностью обмена данными и использования в совместной работе.

Работа выполнена при поддержке РФФИ – проект № 13-07-00218.

ЛИТЕРАТУРА

1. Когаловский М.Р., Калиниченко Л.А. Концептуальное и онтологическое моделирование в информационных системах // Программирование. — 2009. — Т. 35. — №5. — С. 3-25.
2. Uschold M. Ontologies and Database Schema: What's the Difference? 2011. URL: www.slideshare.net/UscholdM/
3. Laallam F.Z., Kherfi M.L., Benslimane S.M. A survey on the complementarity between database and ontologies: principles and research areas// Int. J. Computer Applications in Technology. — 2014. — V. 49. — No 2. — P. 166–187.
4. Зицерман В.Ю., Кобзев Г.А., Фокин Л.Р. Возможности и перспективы информационных технологий в подготовке и распространении справочных данных: свойства веществ и материалов // Научно-техническая информация. Серия 1. — 2004. — №2. — С. 7-14.
5. Degtyarenko K., de Matos P., Ennis M. et al. **ChEBI:** a database and ontology for chemical entities of biological interest// Nucleic Acids Research. — 2008. — V. 36. — Database issue. — D344–D350.
6. Серебряков В.А., Теймуразов К.Б., Хайруллин Р.И., Еркимбаев А.О., Зицерман В.Ю., Кобзев Г.А., Трахтенгерц М.С. Практическая реализация системы интеграции теплофизических данных на основе онтологической модели предметной области // ИНФРАСТРУКТУРА НАУЧНЫХ ИНФОРМАЦИОННЫХ РЕСУРСОВ И СИСТЕМ: Труды Четвертого Всероссийского симпозиума (С.-Петербург. 6-8 октября 2014 г.). — М.: ВЦ РАН, 2014. — С. 415-421.