Фактографические базы данных как компоненты научной информационной системы ♣

© Е.Г.Колесниченко, С.А.Лосев, А.Л.Сергиевская
Институт механики МГУ им. М.В.Ломоносова
Losev@imec.msu.ru, Sergievska@imec.msu.ru

В.В.Варламов, В.В.Чесноков
Научно-исследовательский институт ядерной физики им. Д.В.Скобельцына
МГУ им. М.В.Ломоносова
varlamov@depni.npi.msu.su, chesn@depni.sinp.msu.ru

Аннотация

Рассматривается задача включения баз фактографических данных в состав научной информационной системы. Важнейшая проблема, возникающая при соотнесении имеющихся информационным фондов, - соответствие описания математических моделей предметной области и фактографических данных. Учитывается разнообразие типов данных (литературные, экспериментальные, рекомендуемые данные, и др.), их полнота и непротиворечивость.

Повышение интеллектуальности информационных систем напрямую связано с организацией их взаимодействия с имеющимися и будущими базами данных и знаний, основанных на различных теориях и моделях.

Обсуждается имеющийся опыт разработки фактографических баз данных в области физико-химической газовой динамики, проектирования и создания Web-интерфейса. Приводятся примеры запросов и результатов поиска.

1. Постановка задачи - краткое описание научной информационной системы

Фундаментальной научной проблемой естественнонаучной предметной области (ПО), в частности, физико-химической газодинамики является построение математических моделей, описывающих рассматриваемые явления, а также обеспечение этих моделей фактографической информацией о значениях фигурирующих в этих моделях констант.

Наиболее эффективным средством информационного обеспечения как фундаментальных, так и прикладных исследований в настоящее время являются реляционные базы данных, снабженные мощными и гибкими поисковыми системами.

Создание баз данных по различным характеристикам веществ, используемых в науке и технике, является весьма актуальной задачей. В то же время следует подчеркнуть, что эти характеристики и их значения определяются математической моделью исследуемого процесса или явления, в рамках которой они и используются. В связи с этим не менее актуальной становится и задача создания информационных систем по имеющимся математическим моделям и соотнесении фактографических баз данных соответствующим математическим моделям. Как показывает опыт, решение этой последней задачи требует достижения определенной степени согласия между специалистами относительно ценности различных моделей и их взаимных отношений, а также относительно используемой фактографической информации.

До последнего времени основным путем для достижения такого общепринятого мнения служила организация конференций и семинаров по затрагиваемым вопросам. В настоящее время актуальной представляется задача разработки новых организационных форм совместной работы широкого круга заинтересованных специалистов над решением рассматриваемых проблем (в нашем случае – математического моделирования процессов физико-химической газодинамики). Развитие современных сетевых информационных технологий открывает новые перспективы в решении этой задачи.

Концепция информационной технологии, обеспечивающей информационные потребности научных исследований и приложений в конкретной предметной области, предполагает создание

Труды 5ой Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL2003, Санкт-Петербург, Россия, 2003.

функционирующей в сети Интернет открытой сопровождаемой системы, способной к непрерывному развитию и позволяющей заинтересованному специалисту принять участие в ее создании и сопровождении.

Общая идея проекта заключается в том, что научная информационная система по физико-химической газовой динамике создается параллельно с электронным журналом по этой науке. Доступ ко всем компонентам осуществляется с помощью стандартных браузеров.

Предполагается, что информационная система содержит в первую очередь описание семантической структуры рассматриваемой науки средствами XML с помощью соответствующей схемы.

Такой подход решает три задачи:

- обеспечение широкого обсуждения этой структуры научной общественностью в рамках указанного электронного журнала;
- возможность включать в качестве компонентов рассматриваемой информационной системы появляющиеся в Интернет базы данных и пакеты прикладных программ с адекватным указанием их места в данной науке;
- возможность отслеживать дальнейшее развитие науки путем соответствующей доработки схем или создания альтернативных схем, учитывающих последние достижения.

Наполнение информационной системы осуществляется частично коллективом исполнителей как самостоятельно, так и путем организации ссылок на имеющиеся в Интернет-ресурсы, частично с помощью указанного электронного журнала.

В настоящее время существует достаточно большое число информационных систем, ориентированных на описание отдельных фрагментов физико-химической газовой динамики, начиная от баз данных и кончая коммерческими пакетами прикладных программ для инженерных расчетов. Основной спецификой этих продуктов является их замкнутый характер, не позволяющий в полной мере отобразить дальнейшее развитие науки.

Кроме того, большинство из них ориентированы на локальное функционирование и не доступны по сети Интернет. Идея о совместном сборе библиотеки математических моделей и их фактографического обеспечения, выдвинутая в проекте АВОГАДРО [1], была впоследствии частично реализована в виде баз данных системы АВОГАДРО [2, 3], комплекса KINTVT [4] и компьютеризованного справочника по физико-химическим процессам в газовой динамике [5]. Однако в достаточно полном виде такой библиотеки моделей до настоящего времени не существует.

2. Проблемы - соответствие описания моделей и фактографических данных предметной области

Научное знание можно представить как сложно организованную совокупность интерпретированных формальных систем. Каждая такая формальная система описывает тот или иной материальный (такой как физическая система) или знаковый (такой как компьютерная программа) предмет исследования, поэтому мы будем называть её отдельной формальной теорией или математической моделью. Все они тесно взаимосвязаны между собой. Ввиду быстрого развития научного знания эта совокупность непрерывно расширяется и перестраивается. Каждый предмет исследования может быть с той или иной точностью описан различными математическими моделями и наоборот, практически каждая формальная теория описывает целый класс объектов исследования (т.е. имеет много различных интерпретаций). Математическая модель является экспликацией содержательных представлений данной предметной области в терминах какой-либо из известных математических структур. Поэтому на множестве формальных теорий индуцируются отношения, связывающие данные математические структуры в рамках математического знания. С другой стороны, между ними существуют соотношения, обусловленные структурой моделируемых областей и выражаемые в терминах объектов данной ПО. Эти два типа отношений сложным образом взаимодействуют между собой. Фактографический материал и ограничения целостности, составляющие содержание конкретной базы данных, основаны на интерпретации результатов исследования предметной области в терминах соответствующей теории, как правило, базирующейся на некоторой математической модели.

Для установления однозначной интерпретации баз данных по тем или иным разделам науки, целесообразна их явная привязка к конкретным теориям и математическим моделям. Они являются совокупностью утверждений о значениях констант одной такой модели или некоторого подмножества множества возможных моделей. Таким образом, научные базы данных должны в первую очередь трактоваться как фактографическая интерпретация таких теорий. Существующие базы данных этому требованию не удовлетворяют, поскольку не связаны с явным описанием моделей, интерпретацией которых они являются. Большинство существующих баз данных формировалось либо из авторских коллекций фактографических и литературных данных, либо в результате вторичной обработки большого объема литературных источников. Как правило, эти базы данных не содержат никаких рекомендаций по применению хранящихся в них величин.

3. Опыт разработки баз данных

Обычная практика подготовки решения газодинамической задачи включает в себя, кроме выбора разностной схемы и программирования, поиск и накопление термодинамических данных о компонентах среды, о динамических и кинетических параметрах процессов, протекающих в газовой среде.

При этом, если термодинамическая информация о компонентах рассматриваемой среды достаточно согласована и достоверна (как, например, в системах ИВТАНТЕРМО, АСТРА [6, 7]), то по характеристикам физических и химических процессов почти всегда оказывается невозможной какая-либо априорная оценка достоверности и согласованности данных, выбираемых из различных литературных источников или из кумулятивных баз данных исходной информации.

Значительная часть проблем такого рода может быть решена благодаря базам рекомендуемых данных по характеристикам физико-химических процессов, в которых наряду с со значениями соответствующих величин хранятся оценки достоверности, сформированные путем экспертных обработок и предназначенные для надежной информационной поддержки указанных задач. Специально введенный базовый атрибут – уровень рекомендуемости - отражает степень универсальности данных по отношению к различным приближениям среды, рассматриваемым в конкретных практических задачах.

В настоящее время функционируют две основные базы данных системы АВОГАДРО – базы данных ЧАСТИЦА и ПРОЦЕСС, предназначенные для хранения справочных и рекомендуемых данных, выполнения штатных и произвольных запросов в интерактивном и пакетном режимах.

База данных ЧАСТИЦА. База данных содержит сведения о свойствах частиц (молекулярный вес, потенциал ионизации, энергия сродства к электрону, энтальпия образования, энергии связей, характеристики электронных состояний, колебательные и вращательные постоянные, и др.). Источником информации для этой базы данных являются справочные издания; при отсутствии справочных данных предлагаются результаты экспертных оценок. В настоящее время база данных содержит характеристики более 200 веществ.

База данных ПРОЦЕСС. База данных содержит коэффициенты математических моделей констант скорости конкретных физико-химических процессов (более 2000 процессов) на определенных диапазонах температуры и давления с указанием погрешности и категории рекомендуемости. Атрибут «категория рекомендуемости» характеризует степень обработанности данных. Возможными значениями этого атрибута являются

сд - справочные данные;
рд - рекомендуемые данные;
пд - предварительно-рекомендуемые данные;
ид - информационные данные;
эд - экстраполированные данные;
од - обзорно-оригинальные данные;
1д - первичные данные.

Первичные данные поступают непосредственно из оригинальных авторских публикаций или от экспериментальных установок. Категория "од" присваивается данным, поступающим из обзорных публикаций; каждый такой вклад информации обязательно сопровождается экспертным комментарием. Информационные данные "ид" получаются в результате первичной экспертной обработки данных, имеющих категории "1д" и "од". При выработке данных категории "ид" впервые выполняется первичная экспертная оценка погрешности. Описание погрешности может быть представлено как в числовом, так и в текстовом виде; в особом случае допускается отсутствие оценки погрешности данных категории "ид".

Предварительно-рекомендуемые данные получаются при обработке обзорно-оригинальных или информационных данных; они либо поступают от внешних экспертов, либо получаются при обработке данных, имеющих категорию "од" или "ид". Предварительно-рекомендуемые данные сопровождаются числовыми оценками погрешности, однако эти оценки требуют дополнительных экспертных уточнений и согласования.

Категория "рд" присваивается данным, которые получены в результате тщательной экспертной обработки данных всех остальных категорий. Эти данные поступают от экспертов с обязательной процедурой рецензирования экспертным советом, либо вырабатываются из имеющихся ресурсов.

К экстраполированным данным относятся данные, полученные путем экспертной экстраполяции данных одной из следующих категорий - "рд", "пд" и "ид" - на полный интервал представления аргументов среды в соответствии с заданными параметрами предметной области .

Справочные данные заносятся в хранимые фонды на уровне стандартных данных и не подлежат исправлению в режиме ведения.

Инфологические схемы всех баз данных системы АВОГАДРО построены в виде многоуровневых ациклических ориентированных графов. Базы данных функционируют автономно под управлением стандартной СУБД реляционного типа с языком манипулирования данными SQL, снабжены диалоговыми интерфейсами различного назначения:

4. Включение функционирующих баз данных в научную информационную систему

В настоящее время проводится реорганизация баз данных системы АВОГАДРО для обеспечения свободного доступа к ним из Интернет. С использованием опыта создания реляционных баз данных [8, 9], накопленного в Центре данных фотоядерных экспериментов НИИЯФ МГУ, для этих целей разработан Web-ориентированный интерфейс, серверная часть которого реализована на основе Web-сервера Apache и СУБД MySQL. CGI-сценарии, организующие взаимодействие с клиентской частью, написаны на языке Perl, обеспечивающем высокую переносимость кода, скорость и надежность работы. Технология разделения исполняемого кода и дизайна, использованная при написании скриптов, позволяет менять внешний вид системы без риска нарушить ее работу. Вся клиентская часть, написанная на HTML с использованием CSS, JavaScript и DOM, функционирует в среде Интернет-браузера и не требует дополнительного программного обеспечения.

Web-интерфейс поисковой системы позволяет задать в удобной понятной пользователю форме массу различных критериев выборки (Рис.1), в частности:

- уникальный номер (код) реакции в системе (диапазон кодов);
- класс(подкласс) реакции;
- система атомов процесса и реагентов;
- набор веществ, в который должны входить все реагенты искомой реакции;
- набор веществ, участвующих в реакции;
- фамилия эксперта, написавшего комментарий к реакции.

Диапазоны параметров задаются через дефис, а списки через запятую или пробел. Для удобства пользователя возможно задание максимального числа результатов поиска, выдаваемое на одной странице (Рис.2).

Реализовано также несколько обзорных запросов, позволяющих ознакомиться с возможностями системы и объёмом базы данных. Можно просмотреть все имеющиеся системы атомов процесса и системы атомов реагента, все имеющиеся классы реакций. Предусмотрена также возможность последовательного просмотра информации о реакциях. Для каждой реакции возможно визуально проконтролировать совместимость ее описания в рамках различных моделей с помощью графика, на который нанесены кривые зависимости константы скорости реакции от температуры разными цветами для разных моделей. Этот график, так же как и странички описания реакций, формируется программным обеспечением серверной части системы «на лету». Подробную информацию о каждой модели, использованной для описания реакции, включая математическую формулу, можно получить, выбрав ее имя из списка (Рис.3).

При интегрировании фактографических данных в научную информационную систему были согласованы логические схемы баз данных и семантическая схема разработанного фрагмента создаваемой информационной системы.

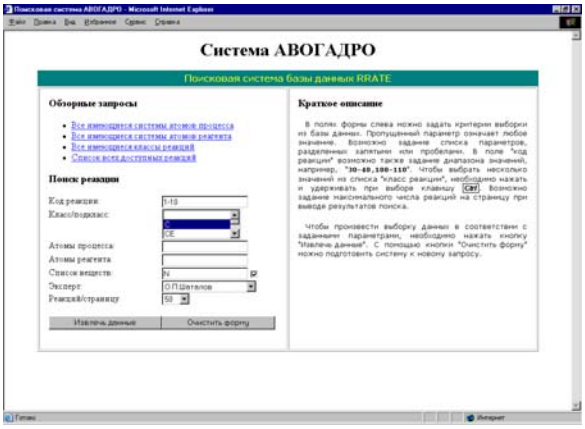

Рис. 1. Поисковая форма реляционной базы данных ПРОЦЕСС системы АВОГАДРО

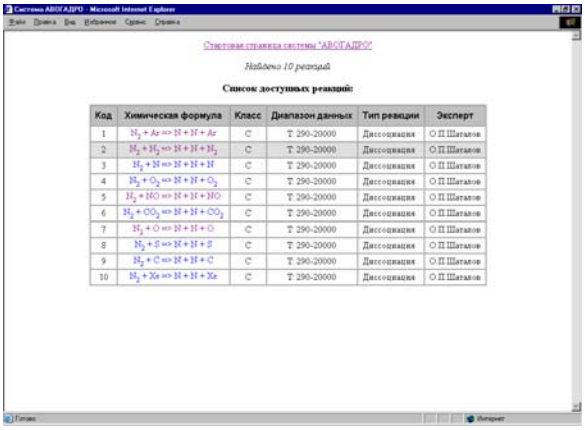

Рис. 2. Форма представления результатов поиска (запрос Рис. 1).

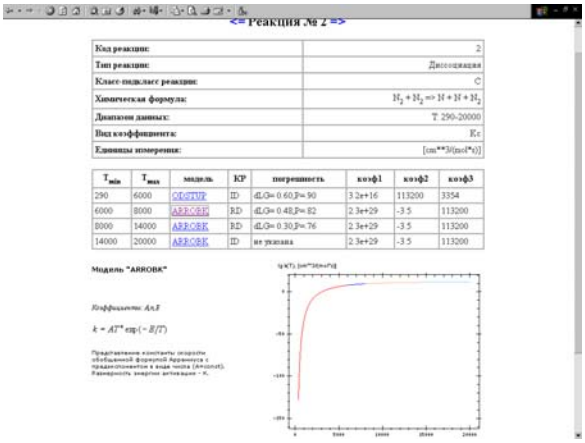

Рис. 3. Данные по конкретной реакции (№2) из списка Рис. 2 (выделена).

5. Перспективы – разработка доступа к разнородным базам данных, развитие баз данных в режиме функционирующей научной системы

Научные базы данных должны с самого начала конструироваться как открытые системы, способные не только расширяться по мере поступления нового материала в рамках данной модели, но и взаимодействовать с другими информационными системами, построенными на основе других моделей. Сами модели, положенные в основу данной информационной системы должны быть описаны с точностью, не допускающей альтернативных толкований. Спецификация моделей и их понятийной базы должна стать неотъемлемой составной частью конкретных информационных систем на основе соответствующих дескриптивных языков. Тем самым однозначно определяется и семантика терминов, использованных в данной системе.

Повышение интеллектуальности информационных систем требует организации их взаимодействия с имеющимися и могущими возникнуть в будущем базами данных и знаний, основанными на других теориях и моделях. Эта задача требует исследования соотношений между различными теориями для достижения интероперабельности соответствующих информационных систем.

Перечисленные задачи приобретают особую важность в свете развития Российской информационной магистрали для науки и образования.

Литература

[1] С.А.Лосев. Система автоматизированного обеспечения физико-химической газодинамики АВОГАДРО: разработка и наполнение. Химия плазмы. Вып.17. М.: Энергоатомиздат. 1993. с. 288-306.

[2] S.A.Losev, E.A.Kovach, A.L.Sergievskaya. Data Base and Computer Experiments in Physico-Chemical Kinetics. 17th CODATA International Conference. Proceedings. Baveno. Italy. October 15-19, 2000.

[3] Э.А.Ковач, С.А.Лосев, А.Л.Сергиевская, Н.А.Храпак. Реализация в системе АВОГАДРО БД “ЧАСТИЦА”. Отчет НИИМ МГУ. N 4405. 1995.

[4] А.Л.Сергиевская, Э.А.Ковач, С.А.Лосев. Опыт информационно-математического моделирования в физико-химической кинетике. М.: Изд-во Моск. ун-та. 1995. 311 с.

[5] Физико-химические процессы в газовой динамике. Справочник. Том 1. Динамика физико-химических процессов в газе и плазме. Под ред. Г.Г.Черного, С.А.Лосева. М.: Изд-во Моск. Ун-та, 1995.

[6] L.V.Gurvich, V.S.Iorish, et.al. IVTANTHERMO - A Thermodynamics Database and Software System for the Personal Computer. User’s Guide. CRC Press, Inc. Boca Raton, 1993.

[7] Б.Г.Трусов. Моделирование химических и фазовых равновесий при высоких температурах (АСТРА-4/РС). М.: МГТУ им. Н.Э.Баумана, 1994.-50с.

[8] I.N.Boboshin, V.V.Chesnokov, E.M.Ivanov, M.E.Stepanov, A.V.Varlamov, V.V.Varlamov. Photon and Charge Particle Reactions and Nuclear Structure Data Bases Upon the MSU INP CDFE Web-site. International Conference on Nuclear Data for Science and Technology. Embracing the Future at the Beginning of the 21st Century (October 7 - 12, 2001). Tsukuba, Japan, Abstracts, Japan Atomic Energy Research Institute, 2001, p. 13-P-1.

[9] И.Н.Бобошин, В.В.Варламов, С.Ю.Комаров, Н.Н.Песков, С.Б.Семин, М.Е.Степанов, В.В.Чесноков. Электронная коллекция научных данных по физике атомных ядер и ядерных реакций ЦДФЭ НИИЯФ МГУ. Труды Четвертой Всероссийской научной конференции RCDL’2002 «Электронные библиотеки: перспективные методы и технологии, электронные коллекции», Дубна, ОИЯИ, 15 – 17 октября 2002 года. Том. 1. ISBN 5-9530-0007-3. ОИЯИ, 2002, стр. 290.

Numerical Databases as Components of Scientific Informaton System

© E.G.Kolesnichenko, S.A.Losev, A.L.Sergievskaya

V.V.Varlamov, V.V.Chesnokov

The problem of inclusion of databases in structure of scientific information system is considered. The major problem arising at correlation available information funds is agreement between the description of mathematical models of a subject domain and the numerical data. A variety of data types (the literary, experimental, recommended data, etc.), their completeness and consistency is taken into account.

Increase of intellectuality information systems directly is connected to the organization of their interaction with available and future databases and the knowledge based on various theories and mathematical models.

Possessed experience development databases on recommended data in the field of physical and chemical gas dynamics, designing and creation of the Web-interface are discussed. Examples of separate queries and results of search are resulted.

♣ Работа выполнена при поддержке гранта Российского Фонда Фундаментальных Исследований № 02-07-90075-в