Проблемы накопления экспериментальной информации в рамках научной информационной системы

А.Л.Сергиевская, Е.Г.Колесниченко, С.А.Лосев
Институт механики МГУ
t-mail: sergievska@inmech.msu.su

1 Вступление. Новые возможности в накоплении и публикации научных результатов с помощью компьютерных информационных сетей

Появление всемирных информационных компьютерных сетей открывает новые возможности для накопления и публикации научных результатов. Одну из таких возможностей мы хотим обсудить в данном докладе. Речь пойдет о публикации и накоплении исчерпывающих протоколов экспериментов (в частности, физических) в рамках научной информационной системы. Коллекцию таких протоколов можно рассматривать как некоторую разновидность электронной библиотеки. Здесь мы хотим обсудить ее организацию и структуру.

2 Постановка задачи сбора первичной информации для хранения в электронной коллекции

Традиционные методы публикации научных результатов, обусловленные в большой мере экономическими соображениями, обладают целым рядом хорошо известных недостатков. Одним из них является существенная потеря приобретенной в экспериментальных исследованиях научной информации при публикации результатов исследования, особенно в тех науках, где для получения конечного результата приходится существенно перерабатывать первичные данные эксперимента. Примером такой науки может служить физико-химическая газодинамика.

2.1 Мотивировка - потеря информации при традиционной форме публикации

Поясним существо рассматриваемой проблемы на примере типичного эксперимента в физико-химической газодинамике. Для этого будем использовать следующие понятия. Математическую модель исследуемой системы можно определить как совокупность атрибутов, которые в физике называются наблюдаемыми, состояния и закона эволюции системы. Обычно принимается, что наблюдаемые образуют алгебру над полем действительных чисел. Описанную в языке прикладного исчисления предикатов теорию данной системы мы отождествляем с математической моделью. Эта теория может содержать в своей сигнатуре ряд констант, которые мы иногда будем называть параметрами математической модели.

Кратко эксперимент можно описать следующим образом. Обычно он заключается в том, что на основе измерений значений тех или иных наблюдаемых исследуемой системы, проведенных с помощью некоторых измерительных приборов, определяются значения тех или иных параметров математической модели, предположительно описывающей исследуемое явление. При этом возникают проблемы, обусловленные неединственностью авторских способов обработки первичных данных, множественностью математических моделей исследуемых физических систем, некорректностью обратной задачи определения параметров математической модели из экспериментальных данных.

- Неединственность авторских способов обработки первичных данных.

  Прежде всего отметим, что получение значений наблюдаемых предполагает некоторую предварительную обработку полученных в измерениях данных, опирающуюся как на методы математической статистики, так и на теорию используемых приборов. Выбор адекватного статистического метода, а также способов интерполяции и экстраполяции измеренных значений не единственен. Кроме того, теория прибора, опирающаяся на его математическую модель, тоже не является единственной. Обычно при получении из экспериментальных данных значений некоторых констант данной модели приходится использовать взятые из литературы значения других констант. Поскольку последние тоже определяются с ошибкой и косвенными методами, это вносит дополнительную неопределенность в получаемые результаты.

- Множественность математических моделей исследуемых физических систем.

  Уже сама постановка любого эксперимента опирается на некоторую математическую модель исследуемой системы, которая зачастую считается сама собой разумеющейся. Между тем эта математическая модель, на которой основывается интерпретация проводимого эксперимента, никоим образом не может считаться единственно возможной.

Первая Всероссийская научная конференция
ЭЛЕКТРОННЫЕ БИБЛИОТЕКИ:
ПЕРСПЕКТИВНЫЕ МЕТОДЫ И ТЕХНОЛОГИИ,
ЭЛЕКТРОННЫЕ КОЛЛЕКЦИИ
19 - 21 октября 1999 г., Санкт-Петербург

- Некорректность обратной задачи определения параметров математической модели из экспериментальных данных.

  Наконец, задача определения параметров модели из экспериментальных данных обычно является некорректной с математической точки зрения. Тем самым окончательный результат зависит от выбранного авторами (зачастую неявным образом) способа регуляризации данной задачи. Указанное обстоятельство является еще одним источником неопределенности в получаемых результатах.

Мы привели основные причины неоднозначности получающихся в результате эксперимента окончательных результатов. Кратко можно сказать, что эта неоднозначность обусловлена неединственностью математической модели (теории) исследуемого явления, а приведенные выше соображения можно рассматривать как предварительную расшифровку этого утверждения. В традиционной журнальной публикации принято приводить полученные в результате обработки экспериментальных данных значения параметров выбранной авторами математической модели вместе с дополнительной информацией, позволяющей читателю более или менее полно воспроизвести условия эксперимента. При этом сама модель часто даже не указывается явно. Между тем эта модель вполне может оказаться неадекватной, но при этом приводимых в публикации данных о проведенном эксперименте практически никогда не бывает достаточно для его переобработки с помощью другой математической модели. Тем самым весьма ценная часть полученной авторами статьи информации оказывается утраченной для научной общественности. Учитывая высокую стоимость физического эксперимента и непреходящую ценность экспериментальной информации эту ситуацию следует признать неудовлетворительной.

2.2 . Электронная коллекция протоколов физических экспериментов (база экспериментальной информации) как компонента научной информационной системы

Возможности сетевых информационных технологий позволяют по новому подойти к этой проблеме и поставить вопрос о публикации исчерпывающих протоколов физических экспериментов. Это прежде всего предполагает публикацию всего набора измеренных данных, а также методов их первичной обработки. Кроме того, необходимым условием однозначной идентификации экспериментальных данных является явное указание математических моделей, использованных при их интерпретации. Всего этого, однако, трудно достичь в журнальной публикации в силу ограничений на объем статьи.

Мы полагаем, что решение поставленной проблемы следует искать на пути создания научных информационных систем по различным отраслям знаний и включения библиотеки протоколов экспериментов в эту систему как одной из их компонент. При этом возникает целый ряд вопросов об организации (архитектуре) такой библиотеки и включающей ее информационной системы в целом.

2.3 Предметная область научной информационной системы

При традиционной трактовке семантики базы данных принято считать, что предметной областью базы данных является совокупность материальных объектов, данные об атрибутах которой содержатся в этой базе. При научном подходе в такую трактовку приходится ввести некоторые коррективы.

Рассматривая описание данной материальной области как некоторую теорию, которую при необходимости можно полностью формализовать, мы должны констатировать, что фактическим содержанием базы данных являются параметры этой теории. Если учесть, что таких теорий может быть несколько, мы приходим к выводу, что предметной областью достаточно полной научной информационной системы оказывается не набор материальных объектов, а совокупность соответствующих теорий этих объектов. Особенно наглядным это утверждение становится в развитых физико-математических науках, таких, например, как физико-химическая газодинамика. Мы трактуем эту науку как совокупность математических моделей (формальных теорий) материальных систем, исследуемых в механике сплошных сред.

Хорошо известно, что для каждой материальной системы можно сформулировать много логически неэквивалентных математических моделей, с достаточной точностью описывающих эту систему. Каждый набор параметров, который мы приписываем данной системе, имеет смысл только в рамках некоторой конкретной теории и может трактоваться как набор констант соответствующей формальной теории. И наоборот, каждой конкретной теории соответствует совокупность наборов параметров, отвечающих различным материальным системам, описываемым этой теорией. Поэтому мы приходим к следующей картине. Каждой конкретной материальной системе соответствует некоторое множество математических моделей с определенными значениями параметров, а каждой общей математической модели соответствует множество различных наборов параметров, определяющих конкретные материальные системы, описываемые этой моделью.

С такой трактовкой предметной области научной информационной системы связан ряд общих семантических проблем.

Прежде всего возникает проблема описательной идентификации физической системы. Если нас интересует конкретная материальная система, то неясно, как следует ее идентифицировать в рамках достаточно полной информационной системы, содержащей описанные выше множества данных. Очевидно, что она соответствует некоторому подмножеству конкретных математических моделей. Это обстоятельство должно быть учтено при конструировании такой информационной системы.

Другим проявлением этой трудности является вопрос о том, как задать семантику математической модели, то есть указать материальную систему, которую описывает эта модель.

Таким образом, мы приходим к тому, что фактическим содержанием научной информационной системы является совокупность математических моделей исследуемых в данной науке систем, между которыми существуют достаточно сложные соотношения, часть из которых упомянута выше. Создание такой информационной системы требует выявления этих соотношений и их реализации в рамках этой информационной системы. В связи с этим возникает проблема структурирования предметной области рассматриваемой информационной системы .

3 Возможные пути решения поставленных проблем

Решение поставленных выше проблем мы видим в создании научных информационных систем по различным отраслям знания. В частности, в Институте механики МГУ в течении многих лет разрабатывался проект создания информационной системы по физико-химической газодинамике АВОГАДРО. В этом проекте был апробирован ряд идей о структуре и характеристиках научной информационной системы. Поэтому в дальнейшем изложении мы будем излагать наши представления о научных информационных системах опираясь на опыт создания системы АВОГАДРО и иллюстрируя наши соображения решениями, принятыми в этой системе.

3.1 Опыт создания научной информационной системы по физико-химической газодинамике АВОГАДРО

Целью системы АВОГАДРО являлось информационное обеспечение проведения математического моделирования изучаемых в физико-химической газодинамике явлений. Система создавалась как совокупность следующих компонент: базы математических моделей, базы рекомендуемых данных и базы исходной информации, а также набора вспомогательных инструментальных средств для работы с этими базами.

3.2 Содержание научной информационной системы

Повторим, что в первом приближении мы рассматриваем науку как сложно организованную совокупность математических моделей, отождествляемых с формальными теориями в языке прикладного исчисления предикатов в том смысле, что каждую из имеющихся в литературе математическую модель можно представить в виде аксиоматической теории в указанном языке. В соответствии с этими представлениями научная информационная система должна содержать описание этих теорий в качестве одной из основных компонент.

В системе АВОГАДРО эта компонента была названа базой моделей. Другая компонента научной информационной системы должна содержать информацию о результатах экспериментальных исследований. В системе АВОГАДРО была названа базой исходной информации.

3.3 Методика описания физического эксперимента

Рассматривая базу экспериментальной информации как базу данных по протоколам физических экспериментов естественно поставить вопрос об выборе адекватного набора атрибутов таких протоколов. Он определяется принятой в данной информационной системе методикой описания эксперимента. В общем можно сказать, что такое описание должно содержать определение исследуемой физической системы, указание измеряемых наблюдаемых системы, методики обработки полученных в измерениях значений, а также математических моделей, используемых для интерпретации эксперимента и определяемых в эксперименте параметров этих моделей. Рассмотрим отдельные компоненты этого описания несколько подробнее.

1. Определение исследуемой физической системы.

   В физико-химической газодинамике объектами исследования являются течения конкретных газовых смесей. Точное определение типа течения сводится фактически к фиксации общей математической модели, используемой для его описания. Если такая модель единственная, то конкретизация исследуемой физической системы заключается в задании значений констант этой модели. ли Если же таких моделей несколько, то возникает упомянутая выше трудность с однозначной идентификацией физической системы.

2. Тип экспериментальной установки.

   В большинстве случаев экспериментальные исследования проводятся на некоторых типичных течениях, реализуемых в типичных экспериментальных установках, таких, например, как ударные трубы. Описание таких установок тоже должно входить в описание эксперимента.

3. Методика эксперимента.

   Под методикой эксперимента мы понимаем выбор тех наблюдаемых, значения которых измеряются в эксперименте, а также способа их измерений, т.е. используемых для этого приборов. Существует достаточно традиционный набор методов измерений, таких, например, как интерферометрические, спектроскопические и т.д.. Их описания тоже желательно иметь в рамках научной информационной системы.

   /item Методика обработки результатов экспериментальных измерений.

   Результаты измерений несут в себе обычно ряд погрешностей. Для исключения этих погрешностей существуют различные статистические методы. Использованные методы должны быть явно указаны при описании эксперимента. Кроме того, параметры модели обычно достаточно сложно связаны с набором измеренных значений наблюдаемых. Определение этих параметров требует проведения весьма сложных расчетов, основанных к тому же на ряде допущений. Методика таких расчетов и сделанные допущения тоже должны быть указаны явно. Следует отметить, что эти расчеты обычно базируются на используемой математической модели и , кроме того, часто используют теории измерительных приборов, употребляемых для определения значений наблюдаемых.

Из сказанного видно, что все эксперименты имеют много общих черт, которые могут быть описаны независимо от конкретных экспериментальных исследований. Тогда при описании протокола конкретного эксперимента в рамках научной информационной системы достаточно будет только указать соответствующие ссылки на эти описания.

3.4 Соотнесение экспериментального исследования с математическими моделями

Как уже отмечалось, частичное определение математической модели исследуемого течения происходит уже при определении исследуемой системы, поскольку для этого

определения используется язык соответствующей математической модели. Это, однако, не всегда бывает так, и окончательное определение математической модели происходит при описании методики обработки результатов измерений и допущений, делаемых при такой обработке.

4 Метаданные описания физического эксперимента

Все перечисленные выше характеристики экспериментального исследования можно трактовать как метаданные этого исследования или публикации о нем. К ним же следует отнести библиографические данные этой публикации, а также некоторые дополнительные ее характеристики, например, экспертную оценку достоверности полученных в результате значений параметров.

5 Шаблоны представления экспериментальных данных для введения в информационную систему

Первоначально в АВОГАДРО планировалось создать по литературным публикациям базу данных о результатах экспериментальных исследований (базу исходной информации - БИИ). Предполагалось, что эта работа будет выполнена коллективом экспертов в данной науке. В связи с этим нами была предпринята попытка разработать стандартный формат для описания отдельной публикации - бланк, заполняемый экспертом научной публикации или автором эксперимента.

Бланк для БИИ содержит:

- библиографические сведения об источнике информации;
- сведения о физико-химических данных;
- сведения об условиях получения фактографических данных (указания на экспериментальные установки, методы и условия измерения, методы обработки данных и т.п.);
- описание математических моделей рассматриваемых процессов (набор переменных, вид уравнений);
- комментарии авторов и экспертов, подготавливающих материал для ввода.

Как показал опыт подготовки рекомендуемых данных, содержащихся в базе рекомендуемых данных системы АВОГАДРО, эта информация необходима для выработки обоснованных рекомендаций.

Схема разработанного бланка исходной информации отображала принятый в АВОГАДРО способ фактографического описания данных для одного определенного параметра, приведенного в конкретной работе (публикации). Именно эти величины и представляют наибольший интерес для системы АВОГАДРО. Для них введен специальный термин: Целевой Информационный Элемент (ЦИЭ).

С целевым информационным элементом связано описание конкретного физико-химического процесса, в котором изучался этот ЦИЭ. Такой процесс назван нами Целевым Информационным Процессом (ЦИП). В некоторых публикациях возможно исследование нескольких различных кинетических коэффициентов. Тогда для каждого ЦИЭ в системе АВОГАДРО оформляется отдельный вклад: библиографическая часть таких вкладов будет одинаковой, а распознавание вклада производится по смыслу ЦИЭ.

В БИИ были определены два типа представления информации, приводимой в отдельном бланке: жесткое (заранее структурированное представление данных) и свободное представление с обязательным выделением заранее оговоренных структурных единиц (таких, например, как ЦИЭ и соответствующий ему ЦИП). Жесткий тип описания соответствует установленному распределению полей в форме ввода (экранной и бумажной) с необходимым условием обязательного заполнения всех полей. Для жесткого представления были разработаны списки установленных значений для тех полей, где это имело смысл, специальный синтаксис строчного представления электронных состояний частиц, описания возбужденных состояний частиц, участвующих в процессе, списки допустимых значений для числовых полей, и т.д.

В состав раздела "качественные атрибуты"жесткого представления информации входит

- краткое описание постановки исследования: физические предположения, допущения, ограничения, цели исследования (например, "поступательно-вращательное равновесие среди молекул", если речь в публикации идет о равенстве газовой температуры и температуры вращательных степеней свободы);
- указание на тип установки, которая использовалась для исследования;
- тип и количество способов воздействия на среду;
- начальный состав газовой смеси и компоненты газа, добавляемые в процессе проведения эксперимента;
- измеряемые величины и молекулярные объекты, к которым относится та или иная измеряемая величина, с указанием размерности;
- краткое описание методики эксперимента;
- метод определения ЦИЭ, так например, константу скорости диссоциации молекулы можно определить из полученной в экспериментах зависимости концентрации компонента от времени; хотя методика измерения этой концентрации могла быть и эмиссионной, и абсорбционной, или иной, но во всех случаях метод определения ЦИЭ - один - по измеренной в эксперименте концентрации исследуемых частиц.

Раздел "количественные атрибуты ЦИЭ"исполняет главную роль в фиксации информации о ЦИЭ, полученной из конкретного источника. Основными элементами этого раздела являются имя математической модели и ЦИЭ - функциональная зависимость кинетического коэффициента или динамического параметра от аргументов, введенных при формализации исследуемого физического процесса; при наличии различных моделей ЦИЭ на различных диапазонах аргументов указываются все возможные модели на соответствующих им диапазонах аргументов.

Раздел "оценочные атрибуты"составляют значения коэффициентов математической модели ЦИЭ на выделенных поддиапазонах аргументов с указанием погрешности. Эти значения могут быть дополнены текстом качественного описания погрешности, например, иногда в

публикации приводится только качественное описание поведения погрешности: "погрешность возрастает от одного значения при некоторой температуре до другого значения при другой температуре", но никаких количественных характеристик такого измерения не дается.

Авторские комментарии и экспертные замечания составляют наименее формализованную часть вклада. Они являются произвольными текстами, оценивающими представленные результаты.

6 Организационные аспекты создания базы экспериментальной информации

Как показал опыт разработки и реализации отдельных компонент системы АВОГАДРО создание такой системы силами одного ограниченного коллектива слишком трудоемко. В частности, заполнение разработанных бланков по уже имеющимся публикациям требует огромных затрат труда экспертов. Поэтому мы были вынуждены отказаться от реализации этого пути.

Появление современных сетевых информационных технологий позволяет надеяться, что при наличии указанной выше научной информационной системы, содержащей приведенные выше компоненты описания эксперимента, сделает привлекательной и существенно облегчит для авторов публикацию протоколов результатов их экспериментов в рамках данной системы. Мы полагаем, что такой подход обеспечит плодотворное использование этих результатов в дальнейших научных исследованиях и будет способствовать прогрессу науки.