ОПИСАНИЕ ИНФОРМАЦИОННЫХ РЕСУРСОВ ПО МОЛЕКУЛЯРНОЙ СПЕКТРОСКОПИИ СРЕДСТВАМИ ПЛАТФОРМЫ XML*

А. З. Фазлиев
Институт оптики атмосферы СО РАН, Томск, Россия
e-mail: faz@iao.ru

An approach to create metadata for an information system in molecular spectroscopy is presented. W3C recommendations are specifically used which suppose that metadata are presented as abstracts containing computer processed statements, schemes and ontologies for domain description.

Введение

Данные о спектральных свойствах вещества являются важнейшим источником информации о строении молекул и процессах, происходящих в газовых средах. Спектральные исследования дают уникальную возможность дистанционного изучения состава и физических характеристик изучаемой среды. По этим причинам спектроскопическая информация широко применяется для решения задач астрофизики, атмосферной оптики, физики пламени и ряда других как научных, так и технических проблем.

В настоящее время наблюдается быстрое совершенствование спектроскопических вычислительных методов исследования, ряд авторов, например Теннисон [1], определяют общую ситуацию как "прорыв" в спектроскопии простых молекул, состоящих из 2—5 атомов. Объемы высокоточных спектроскопических данных, которые необходимо обрабатывать, хранить и использовать в различных приложениях, возрастают быстрыми темпами. Как правило, большая часть этих данных распространяется через ftp или пересылкой по почте на твердых носителях. Следствием сложившейся ситуации являются прерывность поступления данных к пользователям и практически полное отсутствие метаданных по причине неосведомленности спектроскопистов в возможностях современных информационных технологий.

В нашей работе дано описание информационных ресурсов (ИР) в области молекулярной спектроскопии, основанное на использовании платформы XML для структурирования данных, описания и обработки метаданных, а также построения базы знаний. Практичес-

*Работа выполнена при финансовой поддержке Российского фонда фундаментальных исследований (гранты № 02-07-90139, № 05-07-90196).

кое использование этих ИР для коллективной работы с разделяемыми данными и машинной обработкой информации на понятийном уровне позволит решить следующие две задачи. С одной стороны, объединить имеющиеся в России данные с международными информационными ресурсами по спектроскопии, а с другой — стать местом сбора результатов научных проектов, выполняемых в России в этой предметной области.

Предыдущие реализации информационных систем по молекулярной спектроскопии [2–4] были ориентированы на работу с человеком, а не на машинную обработку информационных ресурсов. В новой постановке задачи по созданию распределенной информационно-вычислительной системы (РИВС) машинной обработке отдается значительное предпочтение. Это обстоятельство заставляет нас по-иному отнестись к работе с метаданными и поставить задачу по разработке агентов и Web-сервисов. Чтобы создать почву для таких разработок, требуется построение базы знаний в предметной области. Описание подхода к решению этой задачи является целью данной работы.

Информационные системы, ориентированные на представления сервисов, характеризуются тремя уровнями: уровнем данных и вычислений, информационным уровнем и уровнем знаний [5]. В нашей работе детально описаны два первых уровня и перечислены задачи в молекулярной спектроскопии, которые могут решаться в РИВС на уровне знаний.

Ключевым моментом при создании информационно-вычислительной системы (ИВС) является набор задач, решаемых в ней. В ИВС, описанных в работах [2, 3], к их числу относятся обзор параметров спектральных линий из различных источников, получение при заданных температуре и атмосферном давлении частотной диаграммы интенсивностей, частотного профиля коэффициента поглощения и спектров пропускания, конволюция спектра заданной аппаратной функцией, решение прямой спектроскопической задачи, описание структуры молекулы и спектроскопических констант для основного электронного состояния (в том числе потенциальные функции, дипольные моменты, моменты переходов), колебательных и вращательно-колебательных энергий и волновых функций, изотопических эффектов и т. д.

В нашей работе мы упомянем о трех группах задач, представленных в информационно-вычислительной системе (`http://saga.atmos.iao.ru`). Они связаны с задачами о характеристиках молекулы, в первую очередь ее энергии, вычислениями параметров спектральных линий для молекулярных газов и спектральных функций, характеризующих поглощение и пропускание света молекулярными газами.

Для описания генерируемых по запросу пользователя в ИВС html-страниц используется модель документа, содержащего, как правило, структурированные данные, либо извлеченные из баз данных (это относится к экспериментальным данным), либо полученные в расчетах.

Появление в течение последних пяти лет ряда открытых стандартов W3C [6, 7] по описанию структур данных, метаданных и онтологий позволяет существенно изменить ситуацию в этом направлении. Выполненное в работе описание метаданных полностью опирается на эти стандарты.

1. Уровень данных и вычислений

Молекулярная спектроскопия является физической предметной областью, и в ней используются два способа получения данных — эксперимент и расчет. Данные в этой предметной области можно характеризовать тремя информационными структурами, которые связа-

ны с физическими величинами, необходимыми для решения трех групп задач. Значительная часть физических величин, входящих в эти структуры, экспериментально измеряема. Среди них в настоящее время наиболее изученная структура данных — это параметры спектральных линий [8, 9].

Детали структур данных опубликованы в [4, 10]. Приведем общее описание данных в этих группах.

Фундаментальные характеристики молекул

Это характеристики, определяющие энергию молекулы. В зависимости от способа описания ими могут быть либо параметры полного молекулярного гамильтониана (потенциальная энергия, дипольный момент), либо параметры эффективного гамильтониана (вращательные, центробежные и резонансные постоянные, параметры эффективного дипольного момента). К ним необходимо добавить квадрупольные, октупольные моменты молекул и другие параметры, характеризующие межмолекулярное взаимодействие в газах.

Параметры спектральных линий

Это параметры изолированной спектральной линии (интенсивность, центр линии, энергия нижнего уровня, статистический вес верхнего и нижнего состояний, момент перехода и т. д.), параметры идентификации (колебательная и колебательно-вращательная идентификация), а также параметры, обусловленные столкновениями (полуширина, сдвиг давлением, температурная зависимость полуширины и т. д.).

Спектральные функции

К ним относятся коэффициент поглощения, функция пропускания, сечение поглощения и т. д.

Наряду с задачами молекулярной спектроскопии, решение которых можно найти в работах [2–4], в РИВС решаются задачи систематизации спектроскопических констант, дально- и близкодействующей части потенциала, нахождения волновых функций и уровней энергии, задачи определения параметров спектральных линий (центры, интенсивности, полуширины, сдвиги), задачи расчета спектральных функций с полуэмпирическими контурами и т. д.

Для коллективной работы в ИВС пользователю предоставлена возможность самостоятельного формирования структуры массивов спектральных данных и их наполнения конкретными значениями, проведения на их основе расчетов и сравнения с результатами экспериментов. Наша работа в этом направлении состояла, в первую очередь, в создании системы ввода типовых для спектроскопии структур данных.

2. Аннотации информационных ресурсов

Как отмечено во введении, большая часть имеющихся в Интернете информационных систем является документно-ориентированной. В научной вычислительной системе результаты вычислений можно также представлять в виде информационных ресурсов, включающих в себя документы и относящиеся к ним метаданные.

В нашей статье метаданные, с одной стороны, дополняют документ, делая его более информативным для человека, а с другой — позволяют проводить типовую машинную обработку задач предметной области.

Выбор RDF-схемы и онтологий в предметной области определяется задачами, требующими их формирования. Например, для описания документа как абстрактного информационного ресурса ограничиваются форматом Dublin Core (DC), для описания структуры

документа используют формат PRISM и т.д. Эти уровни абстракции лежат достаточно далеко от предметной области.

В настоящее время RDF-схемы и онтологии для молекулярной спектроскопии отсутствуют. Ниже мы кратко остановимся на описании подхода, использованного нами при формировании таких RDF-схем и онтологий на примере задачи о вычислении коэффициента поглощения.

Основой ИВС являются данные и вычисления. Операции с данными проводятся в рамках физической и математической моделей предметной области. В молекулярной спектроскопии существуют два способа получения данных: экспериментальный и расчетный. Для каждого из этих способов сформированы метаданные. Аннотации информационных ресурсов, характеризующие результаты эксперимента, включают в себя описание устройств, условий эксперимента и т.д. Аннотации для результатов вычислений тесно связаны с метаданными, используемыми при описании физической и математической моделей молекулярной спектроскопии. В нашей работе каждый документ обеспечивается двумя типами метаданных: форматным (DC) и предметным. Далее, для краткости, используется термин “аннотация” для обозначения набора нескольких типов метаданных, связанных с документом. Основное назначение аннотации — это создание возможностей для машинного семантического разбора информационных ресурсов научной ИВС.

Появление рекомендаций W3C (RDF и OWL) заставляет переосмыслить роль метаданных в информационных системах. Цель, преследуемая создателями рекомендаций, состояла в предоставлении инструментов и сервисов для разработчиков информационных систем с целью проектирования и реализации высококачественных, значимых, корректных, минимально избыточных и хорошо аксиоматизированных онтологий [6, 7]. На базе создаваемых онтологий решается ключевая задача — создание аннотаций для информационных ресурсов в Интернете.

В молекулярной спектроскопии нами выделено три механизма автоматического аннотирования предметного содержания ресурсов в ИВС для структурированных данных. Первый из них связан с вводом предметных данных пользователем и заведением источника данных. При вводе данных пользователь заносит ту часть метаданных, которая не может быть механически внесена средствами ИВС, а прочие утверждения формируются динамически (например, число записей в источнике данных, их объем и т. д.). Второй механизм обусловлен процессом манипуляций с данными, при котором пользователь создает новые источники данных на основе уже имеющихся в ИВС. Эти операции, описанные в предыдущем параграфе, протоколируются в аннотации к создаваемому пользователем источнику данных. Отметим, что протоколированию подлежат только манипуляции с данными, доступными всем пользователям ИВС. Третий механизм аннотирования связан с решаемыми пользователем задачами. Выходной документ, получаемый после решения задачи пользователем, содержит аннотацию, включающую в себя RDF-описание данных задачи, методов решения, результатов решения и т. д. Необходимыми компонентами к аннотациям являются онтологии задач, предметной области [5], качества и значений. Множество аннотаций, получаемых в результате ввода данных и решения задач, составляет базу знаний молекулярной спектроскопии. Здесь под базой знаний понимается множество утверждений на формальном языке (RDF).

Обязательной для каждого ресурса ИВС, в том числе для неструктурированных данных, является аннотация, построенная по схеме DC. Для удобства пользователя аннотация источника экспериментальных данных содержит ссылку на XML-документ, в котором хранятся эти данные.

3. Формирование и отображение аннотации

Рассмотрим поэтапную организацию работы с метаданными в распределенной системе "Молекулярная спектроскопия" на примере расчета коэффициента поглощения: от создания документа и его метаданных до механизма обмена метаданными.

Физическая и математическая модели, используемые для расчета коэффициента поглощения, определяют следующие функции для программной реализации:

— ввод и регистрацию данных по коэффициенту поглощения;

— обеспечение средств для расчета коэффициента поглощения с возможностью сравнения полученных результатов с результатами других источников;

— хранение информационных ресурсов в формате XML, а их аннотаций — в виде RDF-описаний в общем реестре;

— обмен аннотациями по коэффициенту поглощения в распределенной информационно-вычислительной системе "Молекулярная спектроскопия" (рис. 1).

В РИВС "Молекулярная спектроскопия" созданы следующие модули работы с информационными ресурсами по коэффициенту поглощения:

— модуль формирования данных и метаданных по коэффициенту поглощения;

— модуль расчета коэффициента поглощения;

— модуль сравнения рассчитанных данных с результатами других источников;

— модуль отображения данных;

— модуль отображения метаданных;

— модуль поддержки обновлений аннотаций в РИВС;

— сборщик "мусора" в реестре аннотаций и базе данных;

— модуль обмена аннотациями в РИВС.

Работа с метаданными по коэффициенту поглощения для экспериментальных данных состоит из двух этапов. Первый этап включает в себя формирование метаданных на основе опубликованного описания эксперимента. Метаданные формируются с помощью

Annotation DublinCore XML

Substance	
Absorbing gas	CO2
Broadening gas	self

Data array	
Wave number (number of values, unit) (cm^{-1})	15
Absorption coefficient (number of values, unit) (am^{-2} cm^{-1}, exp)	15
Errors	Yes

Thermodynamical Conditions	
Temperature (°K)	296
Pressure (atm)	0.25
Broadening gas pressure (atm)	0.25

Spectral parameters	
Spectral resolution (cm^{-1})	0.6
Path length (m)	10
Frequency range (cm^{-1})	2397-2576

Reference	
Authors	Winters B.H., Silverman S., Benedict W.S.
Title	Line shape in the wing beyond the band head of the 4.3mkm band of CO2
Journal	JQSRT 1964, v.4, p. 527-537
Commentary	

Рис. 1. Визуализация утверждений, сформированных при вводе экспериментальной информации о коэффициенте поглощения.

Web-формы, показанной на рис. 2,*а*, и сохраняются в таблице базы данных для коэффициента поглощения. Второй этап состоит в формировании метаданных для результатов научного эксперимента (рис. 2,*б*). Запись аннотации в виде RDF-документа проводится на втором этапе.

Проверка корректности вводимых данных выполняется по типовой схеме: данные пользователя преобразуются в XML-документ, и при разборе используется соответствующая XML-схема. После разбора данные заносятся в базу данных.

При формировании RDF-документа по схеме Dublin Core используются данные о пользователе и ряд технических данных, имеющихся в РИВС. Метаданные для ресурса (DC) и метаданные для коэффициента поглощения формируют аннотацию ресурса. Аннотации ресурсов собраны в РИВС в реестре аннотаций.

В РИВС метаданные для коэффициента поглощения, измеренного в эксперименте, формируются вручную, т.е. через формы вносятся пользователем или администратором. Метаданные для коэффициента поглощения, рассчитанного в РИВС, формируются автоматически при работе пользователя в диалоговой системе. К числу метаданных для рассчитанного коэффициента поглощения относятся спектральный диапазон, температура, давление, ограничение на интенсивность линии, тип контура, величина обрывания контура, способ разбиения спектрального интервала при расчете и т.д.

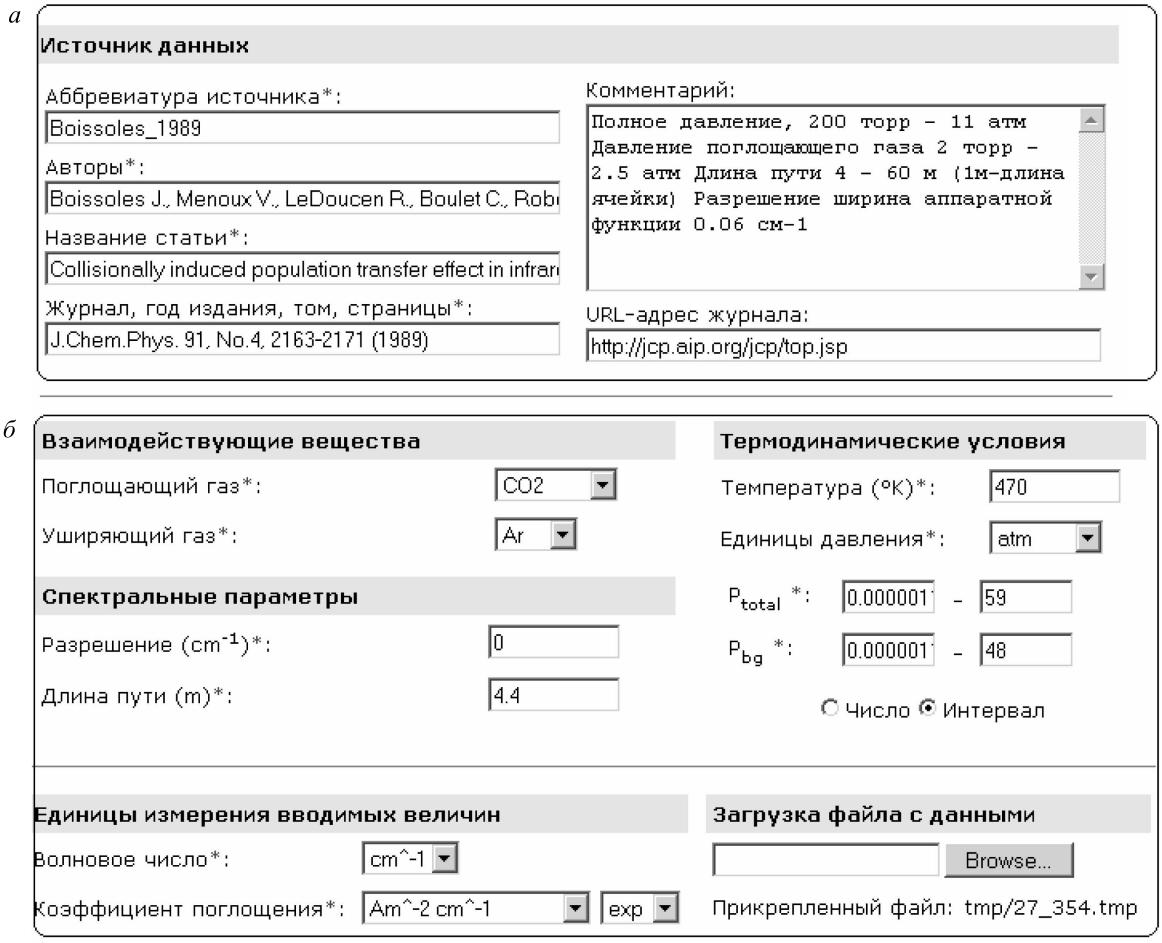

Рис. 2. Формирование источника данных (*а*); формирование метаданных для отдельного массива данных, относящегося к выбранному источнику данных (*б*).

Результат расчета коэффициента поглощения может быть представлен в виде html-страницы, содержащей его табличное или графическое представление, а также ссылки на аннотацию для этого информационного ресурса. Эта часть аннотации отображается вместе со ссылками на метаданные по схеме DC в XML-документ, содержащий данные эксперимента.

Заключение

Описана структура информационных ресурсов распределенной информационно-вычислительной системы "Молекулярная спектроскопия". Предложена структура данных для молекулярной спектроскопии, на примере параметров спектральных линий показан механизм ввода данных пользователя в РИВС, описаны метаданные для коэффициента поглощения и процедура обмена метаданными.

Автор благодарит чл.-корр. РАН С.Д. Творогова, А.Д. Быкова и Б.А. Воронина за консультации и помощь при определении структуры данных в молекулярной спектроскопии и О.Б. Родимову и Н.А. Лаврентьева за подготовку алгоритмов и реализацию программ для расчета коэффициентов поглощения газов.

Список литературы

[1] Tennyson et al. High accuracy ab initio rotation-vibration transitions of water // Science. 2003. Vol. 299. P. 539–542.

[2] Бабиков Ю.Л., Барб А., Головко В.Ф., Тютерев Вл.Г. Интернет-коллекции по молекулярной спектроскопии // Тр. 3-й Всерос. конф. по электронным библиотекам, Петрозаводск, 2001. С. 183–187. (Spectroscopy of Atmospheric Gases `http://spectra.iao.ru`).

[3] Mikhailenko S., Babikov Yu., Tyuterev Vl.G., Barbe A. The databank of ozone spectroscopy on WEB (S&MPO) // Computational Technologies. 2002. Vol. 7. P. 64–70. Spectroscopy & molecular properties of Ozone (`http://ozone.iao.ru`).

[4] Быков А.Д., Воронин Б.А., Козодоев А.В. и др. Информационная система по молекулярной спектроскопии. 1. Работа с данными // Оптика атмосферы и океана. 2004. Т. 17, № 11. С. 921–926.

[5] De Roure D., Jennings N., Shadbolt N. A Future e-Science Infrastructure. Report Commissioned for EPSRC/DTI Core e-Science Programme, 2001. 78 p.

[6] OWL Web Ontology Language Reference, W3C Recommendation 10 February 2004. (`http://www.w3.org/TR/2004/REC-owl-ref-20040210/`)

[7] Resource Description Framework (RDF): Concepts and Abstract Syntax, W3C Recommendation 10 February 2004. (`http://www.w3.org/TR/2004/REC-rdf-concepts-20040210/`)

[8] HITRAN. (`http://cfa-www.harvard.edu/hitran/`)

[9] GEISA. (`http://www.ara.polytechnique.fr`)

[10] Козодоев А.В., Привезенцев А.И., Фазлиев А.З. Организация информационных ресурсов в распределенной информационно-вычислительной системе, ориентированной на решение задач молекулярной спектроскопии // Вычисл. технологии. Спец. выпуск: Тр. IX раб. сов. “El-Pub2004”. Новосибирск, 23–25 сент. 2004 г., Новосибирск, 2005. С. 83–94.

Поступила в редакцию 16 июня 2005 г.