СТРУКТУРИРОВАНИЕ РЕСУРСОВ ИНФОРМАЦИОННОЙ СИСТЕМЫ ПО МОЛЕКУЛЯРНОЙ СПЕКТРОСКОПИИ*

А. Д. Быков, А. В. Козодоев, А. И. Привезенцев, А. З. Фазлиев
Институт оптики атмосферы СО РАН, Томск, Россия
e-mail: faz@iao.ru

Results of variational calculations in molecular spectroscopy increase the spectography data sets in more than one hundred times. It rises a problem how to collect, store and represent this data for Internet users. An hierarchy of the molecular spectroscopy problems used for systematization of data and associated metadata describing the structural and spectral line parameters of molecules is described. Using the proposed hierarchy approach, the data structure is modeled and used for organization of the data uploads. Formalization of the data is implemented using the XML schema. Metadata structure is described with the help of RDF-schema. The problem how to store the information resources consists of two parts: the storage of the elementary data sources and the storage of the complex data sources. The complex data sources are formed according to certain definitive rules, which allow, in particular, to create the data in the Hitran format.

Введение

Работы по созданию информационных ресурсов в области молекулярной спектроскопии атмосферных молекул ведутся в Институте оптики атмосферы (ИОА) СО РАН с начала 80-х годов [1]. Однако качественный скачок в создании информационно-вычислительных систем (ИВС) произошел с появлением персональных компьютеров в начале 90-х, когда была создана система Airsentry [2], имеющая графический интерфейс.

Интернет-технологии позволили сделать следующий шаг в развитии информационно-вычислительных систем коллективного использования по молекулярной спектроскопии. Доступный в сети Интернет информационный ресурс по молекулярной спектроскопии [3] опирался на известные банки спектроскопических данных — HITRAN и GEISA и оригинальные данные [4]. Имеющиеся в этих банках данные определили перечень предметных приложений, доступный пользователю. Отметим, что эти банки данных ориентированы на вычисление спектральных функций. Расширение структур данных

*Работа выполнена при финансовой поддержке Российского фонда фундаментальных исследований (грант № 05-07-90196) и СО РАН (Междисциплинарный проект № 34).

в ИВС, связанных со спектроскопическими исследованиями, было сделано в работе [5] при создании информационно-вычислительной системы S&MO, описывающей свойства молекулы озона. В этой ИВС появились данные о фундаментальных характеристиках изолированной молекулы озона, а именно, уровни энергии, потенциальные и волновые функции и т. д. [6]. С созданием этих систем был выполнен переход от концепции банка данных к концепции информационной системы в области молекулярной спектроскопии. Приложения, соответствующие задачам молекулярной спектроскопии, и существовавшие ранее отдельно от данных, были интегрированы в единую систему с доступом в сети Интернет. Однако в дальнейшем развитие этих информационно-вычислительных систем пошло экстенсивным путем.

С точки зрения подхода, используемого в e-Science [7], модель информационной системы включает в себя три слоя: слой данных и вычислений, информационный слой и слой знаний. Информационный слой ориентирован в первую очередь на обмен информацией между программами. Это означает, что информация должна быть формализована и машинно обрабатываема. Эта задача в рамках подхода Semantic Web решается с помощью языков разметки XML и RDF. Слой знаний создается на основе онтологий, для описания которых W3C рекомендовало язык OWL. Отметим, что в ИВС, описанных в работах [3, 5], был реализован только слой данных и вычислений.

Формирование в ИС информационного слоя потребовало замены концепции данных на концепцию информационного ресурса [8]. В ИВС "Атмосферная спектроскопия" [9, 10] на основе аннотаций как загружаемых данных, так и ресурсов, сформированных в результате решения задач, такой слой был создан. Это потребовало разработки онтологии по молекулярной спектроскопии и онтологии задач молекулярной спектроскопии. Первая упрощенная версия онтологии по молекулярной спектроскопии опубликована в [11].

Этапом для формирования уровня знаний в информационно-вычислительной системе по молекулярной спектроскопии стало создание распределенной информационной системы, в рамках которой должен происходить машинный обмен аннотациями и тем самым должна формироваться база знаний [12].

В настоящей работе описана структура данных и метаданных, используемая при загрузке данных в создаваемую ИВС. В разд. 2 рассмотрена иерархия задач молекулярной спектроскопии, которая стала каркасом для развития инфраструктуры создаваемой нами информационно-вычислительной системы. В разд. 3 представлена схема данных, ориентированная на создание базы данных для хранения спектральных данных, указаны адреса схем для описания данных и метаданных и дано описание онтологии задач, связанных с изолированной молекулой. На примере системы ввода данных об экспериментальных уровнях энергии описаны метаданные, генерируемые приложением. Детали загрузки данных в систему и формирования метаданных рассмотрены в разд. 4. Процесс формирования экспертных ресурсов описан в разд. 5, где рассмотрен механизм реализации сбора, хранения и представления информационных ресурсов в создаваемой ИВС, каждый этап которого связан с изменением структуры информационного ресурса.

1. Иерархия задач

Проектирование информационной системы для предметной области основано на некоторых посылках. В качестве одной их таких посылок мы выбрали возможность разби-

ения предметной области на задачи, что позволило формализовать в информационно-вычислительной системе не только процессы, характерные для молекулярной спектроскопии, но и концепты молекулярной спектроскопии.

Общий подход к классификации задач позволяет в области молекулярной спектроскопии выделить прямые и обратные задачи. Обратные задачи связаны с обработкой данных измерений спектральных функций, что дает возможность в дальнейшем при машинной обработке классифицировать относящиеся к ним данные как экспериментальные.

К элементарным прямым задачам, используемым нами для проектирования информационной системы, относится ряд задач.

1. Задача определения физических характеристик изолированной молекулы (T1). Результатом решения задачи являются вычисленные уровни энергии молекулы, волновые функции, которым соответствуют стационарные состояния и интегралы движения, определяющие квантовые числа для уровней энергии.

2. Задача определения параметров спектральной линии изолированной молекулы (T2). Результатом решения являются частоты переходов (центры линий) и коэффициенты Эйнштейна. Входными данными для задачи являются уровни энергии, волновые функции и квантовые числа.

3. Задача определения параметров контура спектральной линии (T3). Входными данными являются частоты переходов, волновые функции, коэффициенты Эйнштейна и др. Результат решения — вычисленные полуширины, сдвиги, интенсивности, параметры, характеризующие интерференцию спектральных линий, статистические веса.

4. Задача расчета спектральных функций (T4). Входными спектральными данными являются параметры спектральных линий взаимодействующей молекулы. Рассчитываются коэффициенты поглощения, функция пропускания и т. д. при заданных термодинамических и электромагнитных условиях.

5. Измерения спектральных функций (E1). Проводятся измерения спектральных функций. Результатами, значимыми для ИВС, являются значения спектральных функций и метаданные об условиях проведения эксперимента.

Эти задачи образуют иерархию. Например, в простейшем случае для решения задачи T3 необходимо иметь решение задачи T2, или, иными словами, входные данные задачи T3 должны включать в себя выходные данные задачи T2. Выделение первых двух классов обусловлено важным физическим свойством, а именно: свойства изолированных молекул не зависят от термодинамических параметров.

К элементарным обратным задачам относятся:

1. Задача определения параметров спектральной линии взаимодействующей молекулы (ET1). Входными данными являются измеренные спектральные функции и условия измерения. Результат решения задачи — параметры спектральных линий взаимодействующих молекул.

А. Подзадача определения центров спектральных линий (ET1.1). Результатом решения являются частоты переходов (два типа: центры линий, отнесенные к условиям их существования в вакууме, центры линий, отнесенные к конкретным термодинамическим и электромагнитным условиям).

Б. Подзадача определения интенсивностей спектральных линий (ET1.2). Результатом решения являются интенсивности, отнесенные к центрам спектральных линий при заданных термодинамических и электромагнитных условиях.

В. Подзадача определения полуширин, сдвигов и температурных зависимостей полуширин и сдвигов (ET1.3). Результатом решения задачи являются значения параметров контура спектральной линии (полуширина линии, обусловленная столкновениями молекул, сдвиг линии, обусловленный давлением, и температурная зависимость полуширины линии).

Г. Подзадача определения параметров смешения линий (ET1.4).

Д. Подзадача определения коэффициентов Эйнштейна (ET1.5). Результатом являются коэффициенты Эйнштейна, отнесенные к частотам перехода.

2. Задача идентификации спектральных линий (T5). Результатом является установление связи между частотами перехода и квантовыми числами.

3. Задача определения уровней энергии изолированной молекулы (T6). Результатом является список уровней энергии с приписанными к ним квантовыми числами, погрешности определения уровней энергии и число переходов, использованных для определения значения уровня энергии.

Это ключевые задачи. Часть этих задач уже реализована в виде приложений в информационной системе. Для описания предметных ресурсов иерархия задач преобразована в онтологию задач предметной области и используется для формирования базы знаний в молекулярной спектроскопии.

Стоит отметить, что анализ данных, находящихся в базах данных Hitran, JPL и Beamcat, проведенный, например, в [13], показывает, что они относятся только к задачам T2, T3, T4, E1 и E2. Однако надо отметить, что экстенсионал этих баз данных существенно у́же, чем следует из классификации задач. Например, в задаче T3 в базе данных как Hitran, так и Geisa учитываются только самоуширение и уширение воздухом, тогда как в информационной системе, описанной в [9], дополнительно учитывается уширение рядом инертных газов и парами воды.

2. Структура данных и метаданных

В молекулярной спектроскопии изучаются спектры, которыми обладают молекулы. Объектами в молекулярной спектроскопии являются молекулы и излучение. Свойства этих объектов определяют интенсионалы модели данных в информационной модели предметной области. Значения свойств составляют экстенсионал модели данных.

Структура, используемая для хранения данных в ИВС, показана на рис. 1.

Используемые при загрузке схемы данных и связанные с ними метаданные и онтологии можно найти в сети Интернет по адресам:

`http://saga.atmos.iao.ru/data/xsd/tasks/version3/substance/H2O.xsd`;
`http://saga.atmos.iao.ru/saga2/meta/get/v2_T1_for_global.owl`;
`http://atmos.iao.ru/Ontology3/task_t1.owl`;
`http://atmos.iao.ru/Ontology3/Task_T6.owl`.

За основу для построения онтологии задач принят подход, в котором задача является системой, для описания которой используется IPO-модель. Степень детализации входных и выходных данных, а также методов их обработки разная. Так, к числу включенных в рассмотрение метаданных для входных и выходных данных относятся их интенсионалы и ряд атрибутов, характеризующих количественную сторону экстенсионала данных. В выбранной модели метаданные для входных данных представляют собой ссылки на URI ресурсов. Количественные значения, содержащиеся в метадан-

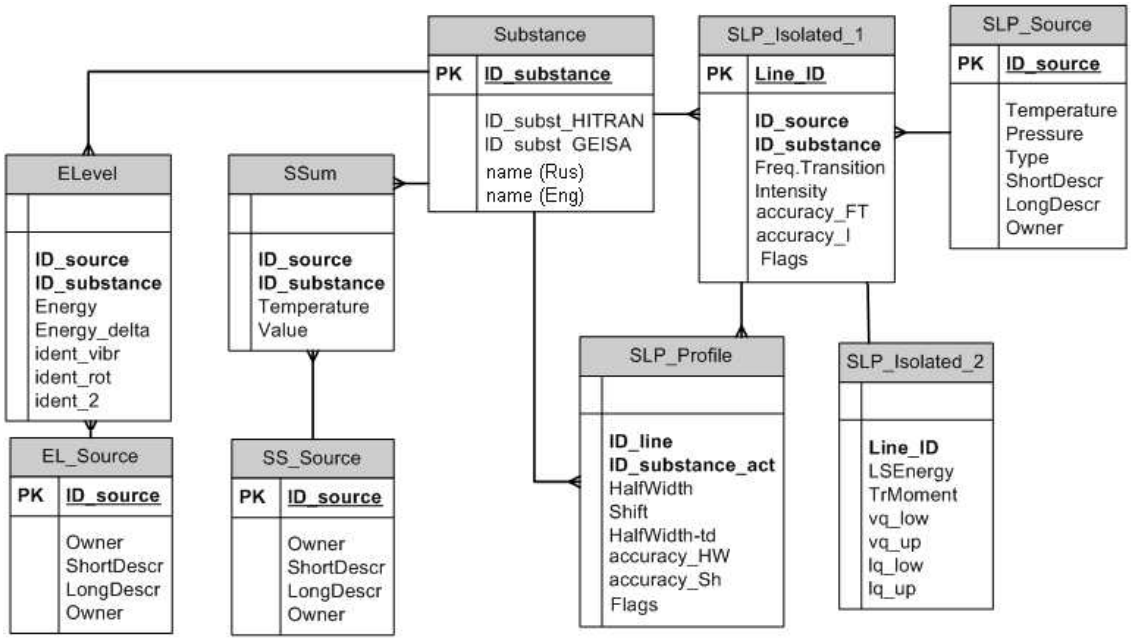

Рис. 1. Схема базы данных, используемой для хранения параметров спектральных линий, уровней энергии и статистических сумм

ных, формируются динамически при загрузке файлов пользователя только для выходных данных задачи. При решении задач в рамках ИВС предполагается формирование метаданных о входных данных также с количественными атрибутами.

Онтология задач T1 и T6, созданная нами, ориентирована на описание метаданных, относящихся к задачам, решения которых загружаются пользователем в ИВС. Метаданные, наряду с интенсионалами входных и выходных данных задачи, описывают некоторые количественные характеристики загруженных данных, например, для задачи T6 число загруженных уровней энергии, их минимальное и максимальное значения, максимальное квантовое число J (полный момент) и т. д.

3. Загрузка данных (уровни энергии молекулы)

Ввод данных в ИВС сформирован в соответствии с иерархией задач. Для каждого класса задач созданы отдельные приложения. Общей для всех процедур ввода данных является процедура создания источника данных, с которым связываются загруженные ресурсы. Источник данных характеризуется уникальным идентификатором, названием, связью с публикацией, хранящейся в базе данных, и классом задач.

В случае отсутствия в библиографической базе данных необходимой публикации пользователь может создать необходимую ему библиографическую ссылку. В системе используется три типа публикаций: статья, монография и Интернет-ссылка. При вводе данных, подготовленных в виде файла, содержащего колонки символов, пользователь с помощью интерфейса описывает интенсионал данных. Загруженный на сервер файл преобразуется в XML-документ и проверяется на соответствие экстенсионала типам данных, описанных в XML-схеме. После разбора XML-документа данные заносятся в базу данных и формируются связанные с ним метаданные. Например, для задачи T6 интенсионал входных данных содержит уровень энергии молекулы, погрешность опре-

Структура метаданных задачи T6

	Интенсионал	Тип	Назначение
1	Пик лист	URI	—
2	Метод	String	—
3	$E_{\min}$	Float	Минимальное значение уровня энергии в массиве данных
4	$E_{\max}$	Float	Максимальное значение уровня энергии в массиве данных
5	N	Integer	Число уровней энергии
6	Угловой момент $J_{\max}$	Integer	Максимальное значение углового момента
7	Тип квантовых чисел	String	(Нормальные моды, BN2, Schwenke)
8	ΔE	Boolean	Признак присутствия погрешностей
9	n	Boolean	Признак присутствия значений числа переходов, использованных для определения уровня энергии

деления уровня энергии, число переходов, использованных для определения уровня энергии, и квантовые числа, характеризующие уровень энергии. Обязательными для загрузки элементами интенсионала являются уровень энергии и хотя бы один набор квантовых чисел. Структура метаданных этой задачи представлена в таблице.

4. Формирование экспертного ресурса

Загрузка пользовательских данных в описываемой ИВС является основным способом наполнения системы новыми элементарными наборами данных. Как экспериментальные, так и расчетные данные, относящиеся к любому из перечисленных выше классов задач, могут попасть в ИВС только в результате их загрузки пользователем или решения соответствующей задачи в информационно-вычислительной системе. При загрузке данных пользователем отсутствует связь загружаемых данных с другими ресурсами, уже находящимися в ИВС. Загружаемые пользователем данные связываются только со своими аннотациями [13] и образуют элементарные ресурсы. Как было описано выше на примере задачи T6, эти ресурсы характеризуются источником данных, содержащим библиографическую ссылку.

Как правило, данные загружаются в ИВС в виде файлов. Структуры данных, используемые в файлах, могут не соответствовать структурам данных, используемым для хранения ресурсов. Отметим, что наиболее распространенной структурой данных, используемой в загрузочном файле, являются колонки, строки с фиксированными позициями данных и деревья, размеченные с помощью языка разметки XML. Существуют и иные способы форматирования спектральных данных в молекулярной спектроскопии [14].

Конкретная структура данных, используемая при загрузке, обусловлена задачей молекулярной спектроскопии, решением которой эти данные являются. Структура данных, применяемая для их хранения, может быть иной и обусловленной задачами, которые их используют.

Рассмотрим на примере задачи формирования составных ресурсов механизм изменения структуры ресурсов. На рис. 2 показана последовательность формирования экспертного информационного ресурса. После загрузки ресурсы имеют статус персо-

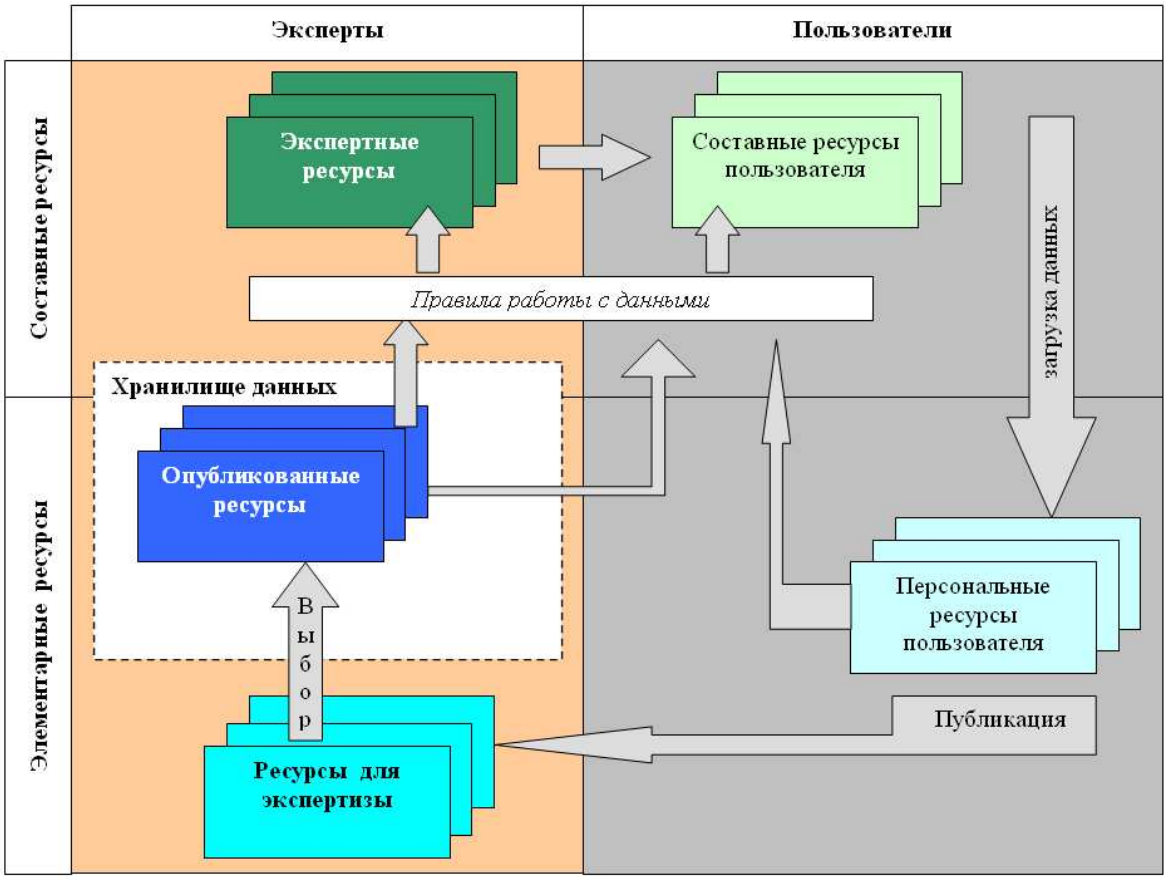

Рис. 2. Схема процессов для работы с данными

нальных ресурсов и доступны только собственнику. Для организации доступа к загруженным пользователем ресурсам используется процедура экспертного отбора. Каждый пользователь РИВС по определенной процедуре может опубликовать свои элементарные ресурсы, загруженные в ИВС, после экспертной оценки. Статус рекомендованного для публикации ресурса означает невозможность его изменения собственником ресурса. Рекомендованные к публикации ресурсы становятся доступными экспертам. Отобранные экспертами ресурсы приобретают статус опубликованных ресурсов. Они помещаются в хранилище данных и становятся общедоступными.

На основе опубликованных ресурсов все пользователи, в том числе и эксперты, могут формировать составные ресурсы в рамках правил, поддерживаемых в ИВС [15]. Правила должны обеспечивать механизм создания составного ресурса. Формирование хранилища данных решает проблему непрозрачности процедуры формирования ресурсов, имеющихся в банках данных Hitran и Geisa. Созданные экспертами составные ресурсы могут также применяться пользователями для их задач. При этом структура данных, предоставляемых пользователю, может формироваться самим пользователем. Для параметров спектральных линий по умолчанию выбран формат файла данных, используемый в банке данных Hitran.

Структура ресурсов, формируемая экспертами, определяется прикладными задачами, для которых эти ресурсы являются входными данными. Она может не совпадать со структурой, используемой для хранения данных в хранилище данных, содержащем опубликованные ресурсы.

Заключение

Представлен подход к структурированию данных и метаданных для создания информационной системы по молекулярной спектроскопии. Подход основан на иерархии задач предметной области. Рассмотрена процедура загрузки данных в ИВС с генерацией количественных метаданных для нескольких задач иерархии. Структура загружаемых данных определяется XML-схемами, а метаданных — RDF-схемами. Связи между интенсионалами данных описаны в онтологии задач предметной области. Таким образом, описание информационных ресурсов для части задач молекулярной спектроскопии соответствует требованиям, согласно которым их можно отнести к ресурсам семантического веба. Предложенное решение задачи формирования экспертного информационного ресурса относится к следующему этапу развития семантического веба, формализация которого в данное время не завершена. В частности, этот этап требует средств описания правил для семантических ресурсов.

В настоящее время структурированы данные и описаны предметные метаданные для задач T1, T4 и T6. В ИВС (`http://saga.atmos.iao.ru`) реализованы ввод данных, генерация метаданных и формирование индивидуалов онтологии задач молекулярной спектроскопии. Проводится работа по структурированию данных и описанию метаданных для оставшихся задач молекулярной спектроскопии.

Список литературы

[1] Войцеховская О.К., Розина А.В., Трифонова Н.Н. Информационная система по спектроскопии высокого разрешения. Новосибирск: Наука, 1988. 150 с.

[2] Golovko V.F., Nikitin A.V., Chursin A.A., Tyuterev V.G. Information system AIRSENTRY for modeling atmospheric IR-spectra and radiation transmission in the atmosphere // Proc. of the 2nd Intern. Workshop ADBIS'95. Vol. 2. M., 1995. P. 12–14.

[3] Бабиков Ю.Л., Barbe A., Головко В.Ф. и др. Интернет-коллекция по молекулярной спектроскопии // Тр. 3-й Всерос. конф. "Электронные библиотеки: перспективные методы и технологии, электронные коллекции". Петрозаводск, 2001. С. 183–187. `http://spectra.iao.ru`

[4] Банк данных CO_2. `ftp://ftp.iao.ru/pub/CDSD-296`, `ftp://ftp.iao.ru/pub/CDSD-1000`

[5] Михайленко С.М., Бабиков Ю.Л., Тютерев В.Г., Barbe A. The databank of ozone spectroscopy on WEB (S&MPO) // Comp. Technologies. 2002. Vol. 7. P. 64–70. `http://ozone.iao.ru`

[6] Тютерев В.Г. Глобальные вариационные и эффективные методы расчетов положений и интенсивностей спектральных линий трехатомных молекул: некоторые тенденции и особенности нового поколения спектроскопических информационных систем // Оптика атмосферы и океана. 2003. Т. 16, № 3. С. 245–255.

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

[8] Когаловский М.Р. Научные коллекции информационных ресурсов в электронных библиотеках // Тр. 1-й Всерос. науч. конф. "Электронные библиотеки: перспективные методы и технологии, электронные коллекции". СПб., 1999. С. 16–31.

[9] Быков А.Д., Воронин Б.А., Козодоев А.В. и др. Информационная система по молекулярной спектроскопии. Ч. 1: Структура информационных ресурсов // Оптика атмосферы и океана. 2004. Т. 17, № 11. С. 816–820. http://saga.atmos.iao.ru

[10] Фазлиев А.З. Описание информационных ресурсов по спектроскопии средствами платформы XML // Вычисл. технологии. 2005. Т. 10. Спецвыпуск. Ч. 1. С. 39–46.

[11] Родимова О.Б., Творогов С.Д., Фазлиев А.З. Онтология молекулярной спектроскопии атмосферных газов // Тр. 5-й Всерос. науч. конф. "Электронные библиотеки: перспективные методы и технологии, электронные коллекции". СПб., 2003. С. 211–215.

[12] Козодоев А.В., Привезенцев А.И., Фазлиев А.З. Аннотирование информационных ресурсов в распределенной информационной системе "Молекулярная спектроскопия" // Тр. 7-й Всерос. науч. конф. "Электронные библиотеки: перспективные методы и технологии, электронные коллекции ". Ярославль, 2005. С. 80–86.

[13] Feast D.G. A Spectral Line Database for Millimeter and Submillimeter Wave Propagation in the Earth's Atmosphere. Research Report N 99-1. Institute of Applied Physics, Bern, 1999.

[14] Lancashire R., Davies T. Spectroscopic data: the quest for a universal format // Chemistry International. 2006. Vol. 28, N 1.
http://www.iupac.org/publications/ci/2006/2801/3_ref5.html

[15] Козодоев А.В., Фазлиев А.З. Информационная система для решения задач молекулярной спектроскопии. Ч. 2: Операции преобразования наборов параметров спектральных линий // Оптика атмосферы и океана. 2005. Т. 18, № 9. С. 760–764.

Поступила в редакцию 11 мая 2007 г.