Онтология по молекулярной спектроскопии атмосферных газов

© О.Б.Родимова, С.Д.Творогов, А.З.Фазлиев

Институт оптики атмосферы СО РАН, Томск
rod@iao.ru, tvorogov@iao.ru, faz@iao.ru

Аннотация

Представлено описание информационных ресурсов в области молекулярной спектроскопии, развиваемых в Институте оптики атмосферы СО РАН. Описан как подход к описанию метаданных для этих ресурсов средствами RDF для машинной обработки метаданных, так и онтология, ориентированная на создание непротиворечивой диалоговой системы и поисковых средств.

1 Введение

Работы по созданию информационных систем (ИВС) по молекулярной спектроскопии ведутся в ИОА СО РАН с начала 80-х годов [1]. Основанные на технологиях свойственным машинам серии ЕС эти информационные системы закончили свое существование с появление персональных компьютеров. Вторым этапом развития ИС по молекулярной спектроскопии явились работы Головко В.Ф. и др. [2] приведшие к созданию информационных систем, работающих на клиентском месте.

Появление интернет - технологий позволило сделать следующий шаг в развитии ИС по молекулярной спектроскопии [3-5]. Доступный в Интернете информационный ресурс (http://spectra.iao.ru) опирается на известные банки спектроскопических данных Hitran и Geisa, и, следовательно, заимствует структуру их данных. Большая часть решаемых в информационно-вычислительной системе SPECTRA задач связана с поиском параметров спектральной линии в базе данных. В работе [4] описаны доступные в Интернете информационные ресурсы по спектроскопии, технический способ построения диалоговой системы и html-описание метаданных использованные в ИВС SPECTRA. Нерешенными вопросами, относящимися к метаданным этой ИВС, остались RDF-описание интернет-ресурсов для ИВС и построение онтологии по молекулярной спектроскопии.

Решение указанных проблем, а также расширение данных БД Hitan и Geisa за счет расчетных данных Партриджа и Швенке [6] по молекуле воды, включающих миллионы спектральных линий, оригинальных данных, полученных в ИОА СО РАН и организация доступа к спектроскопическим программам для расчета были положены в основу гранта, поддержанного РФФИ (грант 02-07-00139) [5].

Специфика информационной системы по спектроскопии состоит в том, что спектральные данные включают в себя миллионы линий, используемых при расчетах физических характеристик атмосферы, требующих деталей понимания процессов поглощения в молекулах. В последнее время увеличился научный интерес к слабым линиям, что в свою очередь привело к расширению объема спектральных данных только для молекулы воды в сотню раз. Наиболее серьезной проблемой, решение которой находится в сфере самой спектроскопии, остается верификация данных включаемых в разные спектральные базы данных.

Стоит отметить, что количество данных по молекулярным спектрам, представляемых в Интернете, растет, прежде всего, за счет публикаций в электронных изданиях, и, следовательно, требуется некоторая осознанная схема работы с такими данными. Создаваемый информационный ресурс требует описания на уровне метаданных. Такое описание информационного ресурса, представляемого в Интернете, необходимо проводить в рамках открытых стандартов установленных W3C. Построение метаданных в рамках подхода, использующего RDF-схему [7], требует создания словаря терминов предметной области и установления связей между ними.

Для работы с информационным ресурсом необходима диалоговая система, построение которой должно учитывать структуру спектральных данных и метаданные, связанные с ними. Использование RDF–схемы для этих целей является очевидным, но описательные средства данной рекомендации ограничены. Расширение возможностей, заложенных в RDF-схеме, проводится в рекомендации Web Ontology Language [8].

В данном докладе описаны информационные ресурсы спроектированной информационно-вычислительной системы “Атмосферная молекулярная спектроскопия” [9] и подход к построению онтологии по молекулярной спектроскопии. Исход-

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

ное назначение создаваемой онтологии ориентировано на построение внутренне непротиворечивой диалоговой системы и в будущем поисковой системы по молекулярной спектроскопии на основе имеющегося информационного ресурса.

2 Структура данных

Появление возможности быстрого обмена данными и простота их занесения в базы данных ставит перед спектроскопистами задачи более детального структурирования данных. К числу таких задач относится ввод в БД неполных данных (например, относительных интенсивностей спектральных линий), указание связей между расчетными данными и экспериментальными на основе которых проводилась подгонка, машинный поиск по типу данных (экспериментальные или расчетные). Эти обстоятельства явились определяющими для ввода расширенного формата данных по сравнению с форматом данных Hitran'а.

Планируемое расширение структуры данных содержит информацию об идентификации высоковозбужденных состояний паров воды, данные измерений проводимых в ИОА СО РАН (данные по высоковозбужденным спектрам воды), экспериментальные спектры с указанием источника данных, данные из электронных источников информации, данные пользователей и т.д.. Расчетные данные по спектрам воды требуют для описании таксономии методов расчета, используемых в спектроскопии. Эта проблема обсуждается в следующем разделе.

Генезис данных в БД Hitran и Geisa таков, что в основу классификационной схемы положена гипотеза об изолированной спектральной линии, из которой следует необходимость выделения параметров (центр линии, интенсивность, классификация состояний и т.д.). Согласно этой гипотезе, спектральные характеристики атмосферных молекул можно построить для разных значений температур и давлений, опираясь на данные, вычисленные или измеренные при некоторой температуре и давлении (при нормальных условиях). При этом для одной и той же спектральной линии существует два набора спектральных параметров: один для интервала температур 70-1000K заданный для T=296K, а второй для T>1000K. Стоит отметить, что данные Hitran и Geisa соответствуют нормальным условиям (T=296K и P=1 атм), а для моделирования высокотемпературных спектров требуются данные из HighTemp.

На рис. 1 представлена схема базы данных, обобщающая результаты работы [3].

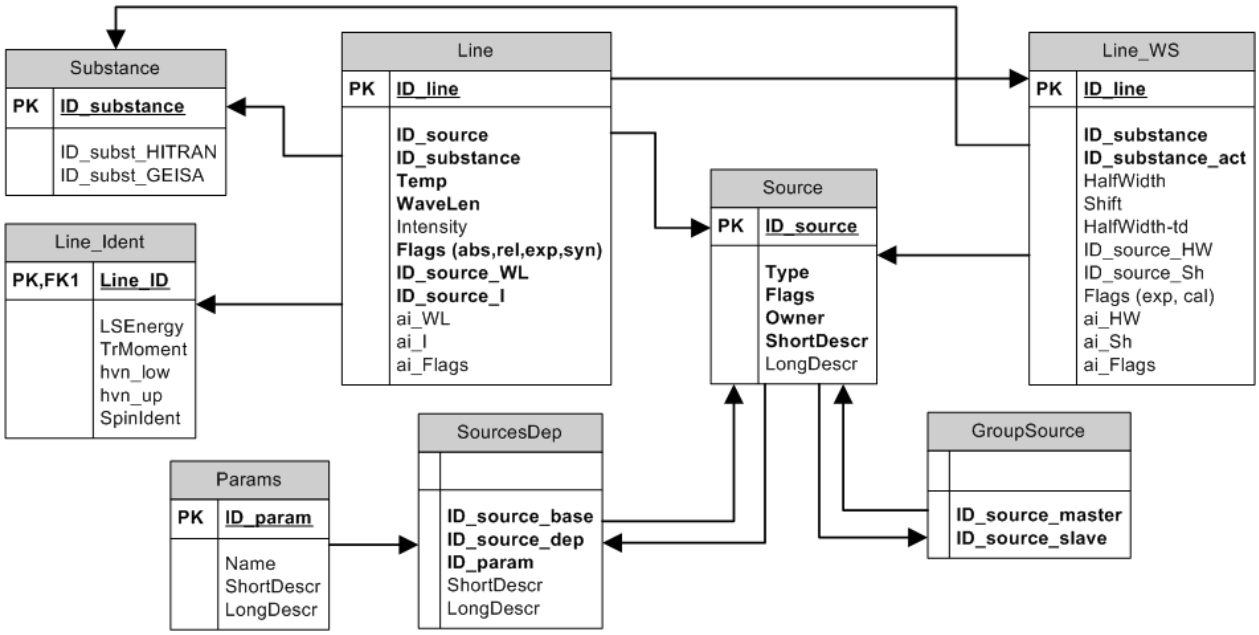

Рис.1.Схема базы данных

При работе с данными, относящимися к характеристикам спектральных линий, обычно используют формат данных, введенный при создании банка спектральных линий Hitran. Менее распространен формат данных использованный создателями банка данных Geisa. Оба формата данных связаны с базовым представлением об основных спектральных свойствах молекулы. Строку данных можно условно разделить на три части. К первой из них отнести данные характерные для изолированной молекулы (центр линии, интенсивность, сила линии, нижний уровень энергии, статистическая сумма и колебательно-вращательная идентификация линии). Ко второй группе данных отнести физические величины, характеризующие процессы взаимодействия молекулы со средой (уширение разными молекулами (в Hitran и Geisa полуширины при самоуширении и уширении воздухом), сдвиг центра линии, температурная зависимость полуширины) и к третьей группе – классы точности и библиографию по

источникам данных. В силу того, что оба указанных банка данных были ориентированы на поддержку решения прикладных задач, такая структура данных отвечала сложившимся потребностям пользователей спектральной информации.

Таким образом, информация о спектральной линии состоит из трех блоков: параметры изолированной линии, параметры неизолированной линии и вспомогательная информация (точности измерения параметров и библиография). Различие между структурой данных в банках Hitran и Geisa состоит в деталях того, как в этих модулях описываются данные.

Перечень основных физических величин представленный в базе данных и используемый в качестве словаря для описания метаданных приведен ниже.

1. Спектральные параметры изолированной молекулы
   - 1.1. Центр линии
   - 1.2. Интенсивность (абсолютная или относительная величина)
   - 1.3. Сила линии
   - 1.4. Нижний уровень энергии
   - 1.5. Статистическая сумма
   - 1.6. Колебательно-вращательная идентификация
2. Спектральные параметры неизолированной молекулы
   - 2.1. Полуширина линии (самоуширение, уширение воздухом, уширение другими молекулами)
   - 2.2. Сдвиг центра линии
   - 2.3. Температурная зависимость полуширины
3. Коэффициенты поглощения (экспериментальные данные)
4. Вспомогательная информация
   - 4.1. Точности (класс точности, абсолютное значение поправки)
   - 4.2. Библиографические ссылки

Представленный список физических величин, характеризующий спектры атмосферных молекул, достаточен для описания информационных ресурсов спектров изолированных молекул. На рис.2 показано какую часть эти понятия составляют в RDF-графе, применяемом для описания соответствующего информационного ресурса.

3 Описание информационных ресурсов

Информационный ресурс, содержащийся в ИВС может быть как статическим, так и динамическим.

Описание статического ресурса, т.е. того который, хотя и размещается в базе данных, но недоступен для изменения пользователю, требует только использования словарных терминов, введенных в предыдущем разделе при описании данных.

Динамический ресурс создается в процессе работы пользователя в ИВС и содержит данные являющиеся результатами расчетов, проводимых по набору алгоритмов. Ясно, что использование разных алгоритмов для вычисления одной и той же физической характеристики будет представлять разный информационный ресурс. Последнее должно быть отражено в метаданных, связанных с ресурсом. Наличие в ИВС вычислительной компоненты позволяет пользователю в зависимости от заданных им термодинамических условий и формы контура спектральной линии вычислять спектральные функции, получать спектры низкого разрешения и т.д. В этом случае описанных в предыдущем разделе основных параметров спектральных линий оказывается недостаточно для корректного описания создаваемого ресурса. Требуется расширение понятийного набора. Такое расширение происходит за счет таких понятий как форма контура линии, термодинамические параметры, аппаратная функция и т.д.

Если статический ресурс, генерируемый из БД можно описать в терминах используемых при формировании базы данных, то для описания динамического информационного ресурса по молекулярной спектроскопии требуется дополнить с одной стороны словарем, содержащим понятия вычислительной математики, а с другой стороны словарем с базовыми понятиям квантовой механики, на основе которой строятся методы расчета спектральных характеристик атмосферы.

Базовыми понятиями для описания ресурсов по спектроскопии являются молекулы (класс “ Substance”) и их классификационные схемы (симметрия и модельные представления), экспериментальные (“Experiment”) и расчетные данные (“Model”), спектр (“Spectrum”) и спектральные функции.

Для описания метаданных статического ресурса построена RDF-схема (словарь). На основе этой схемы проведено описание ресурсов, представляющих собой набор html-страниц, сгенерированных из баз данных и содержащих информацию по каждой из спектральных полос молекул, входящих в Hitran и Geisa. Классы “Substance” и “Spectrum” связанны с описанием статического ресурса.

На рис.2. показан граф содержащий связи между классами и их свойствами. Большая часть прямоугольников (литералов) на рисунке соответствует данным из банка спектральных данных Hitran. Всего на схеме отображено 335 утверждений, 189 ресурсов и 39 литералов. Для визуализации утверждений возникающих при RDF описания использовалась программа IsaViz RDF Editor (http://www.w3.org/2001/11/IsaViz).

4 Построение онтологии

Известно, что онтология включает в себя описание классов, свойств и их экземпляров и используется для понимания предметной области как человеком, так и приложениями которым необходима информация о предметной области. Естественным является построение вводной части в молекулярную спектроскопию по структуре созданной онтологии, что должно помочь пользователям в систематизации знаний сконцентрированных на сайте. Построе-

ние классов и свойств онтологии проведено в рамках OWL DL и в существенной степени опирается на RDFS описание ресурсов по молекулярной спектроскопии. При описании использованы следующие конструкты: характеристики свойств, произвольная кардинальность, аксиомы классов. При построении онтологии в части использования новых классов имеется существенное дополнение. Для организации поиска на основе онтологии введен класс “Процесс” свойственный практически всем онтологиям, ориентированным на описание предметных областей естественных наук. К числу подклассов класса “Процесс” в спектроскопии относятся такие классы как “Поглощение”, ”Флуоресценция”, ”Пропускание”, ”Излучение” и т.д.

Для построения графического и OWL/XML представления использовался редактор IsaViz.

В докладе обсуждается связь диалоговой системы со структурами входящими в онтологическое описание. Последнее актуально в силу того факта,

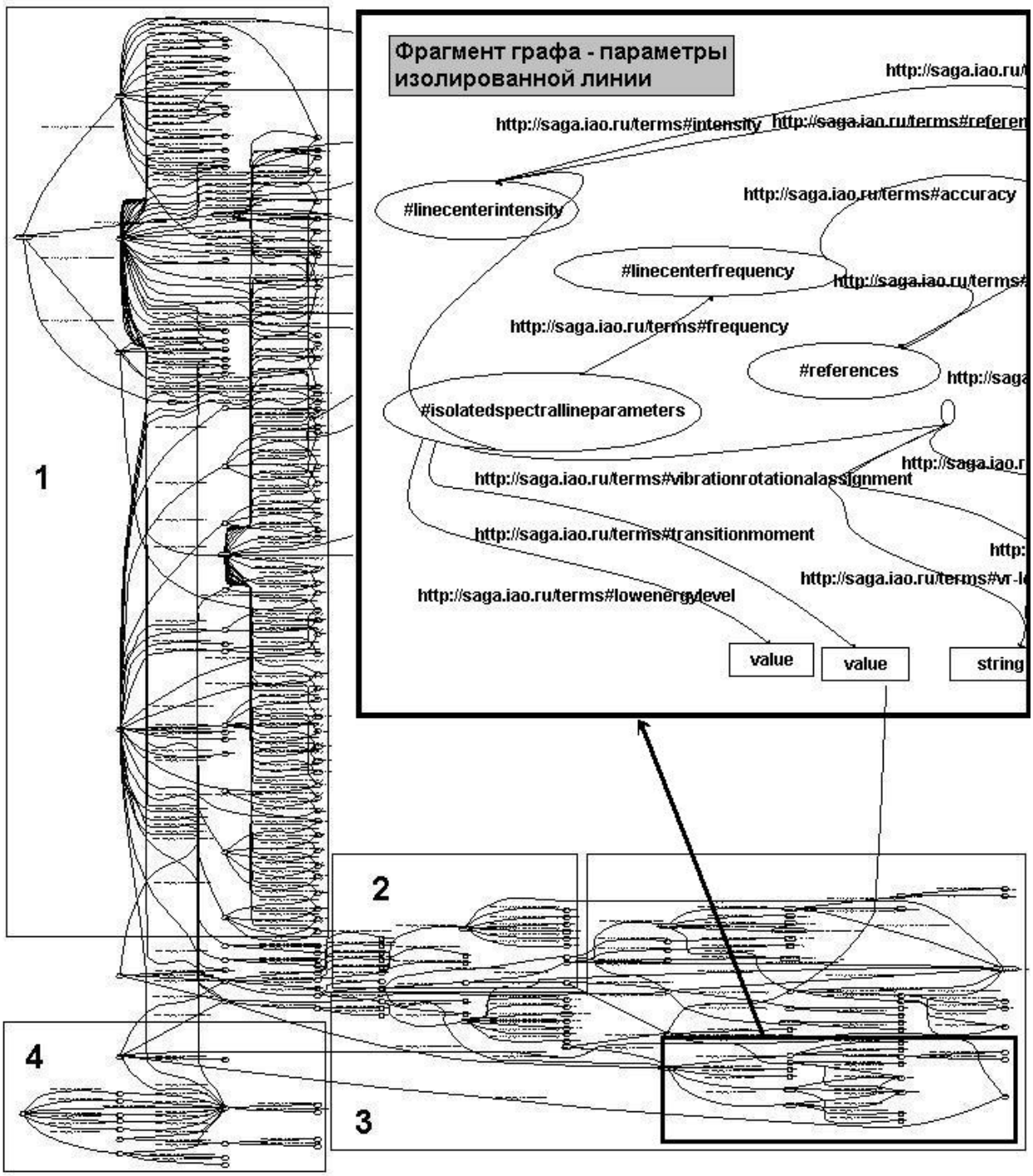

Рис. 2. RDF-граф, описывающий связи между классами ”Спектр” (3), “Спектральные функции” (2) и “Вещество” (1) для атмосферной молекулярной спектроскопии и ориентированный на описание данных из БД Hitran. Объекты помеченные овалами представляют ресурсы, прямоугольниками – литералы и дуги соответствуют свойствам ресурсов.

что научные информационные ресурсы, как правило, являются динамичными в своей структуре [10], что заставляет использовать достаточно гибкие интерфейсы, адаптирующиеся с одной стороны к изменяющимся потребностям пользователей, а с другой стороны, отражающие разные концептуальные подходы, существующие в молекулярной спектроскопии.

5 Структура ИВС

В создаваемой ИС существует три уровня компетентности пользователей (начинающий, прикладник и эксперт). В зависимости от уровня компетентности пользователи имеют разный доступ к данным и разным ветвям дерева меню. Информационная система состоит из трех разделов. В первом – проводится работа с данными. На административном уровне осуществляется ввод данных и контроль их целостности, а на пользовательском – только ввод данных. Можно вводить два набора данных – параметры линий (минимальный набор – центры линий) и спектральные функции (коэффициент поглощения, функцию пропускания и т.д.). Предусмотрено сравнение всех предметно совместимых наборов данных.

Во втором разделе проводится работа с данными по изолированной молекуле. В зависимости от уровня компетентности пользователь может проводить сравнение диаграмм интенсивностей из разных источников данных для одной молекулы, в разрезе полоса или спектральный интервал. Здесь же собраны данные по сечению поглощения молекул и классификация молекул по разным критериям (группам симметрии и …..).

Третий раздел является основным для описания свойств молекулы взаимодействующей с окружением. В этом разделе доступны вычисления спектральных функций (коэффициенты поглощения и функции пропускания). Отметим, что термодинамические условия задаются пользователем до его обращения к разделам ИС в которых проводятся расчеты. Связано это с тем, что в зависимости от них с одной стороны проводится выборка данных только из выбранного пользователем интервала температур, а варианты формы контура линии предлагаются в зависимости от интервала давлений интересующего пользователя.

Технической основой для создания сайта явилось middleware, разработанное в институте оптики атмосферы СО РАН [11].

Литература

[1] Войцеховская О.К., Макушкин Ю.С., и др., Тезисы докладов 6 Всесоюзного симпозиума по молекулярной спектроскопии высокого и сверхвысокого разрешения, Томск, 1982, ч.2., с. 42-44.

[2] Golovko V.F., etc, Information system AIR-SENTRY for modeling atmospheric IR-spectra and radiation transmission in the atmosphere, ADBIS'95 Proc. The 2-nd Int. Workshop, v.2, Moscow, 1995, p.12-14.

[3] Babikov Yu.L., etc, WEB information system: atmospheric spectroscopy, Proc. SPIE "7-th Int. Symp. on Atmos. and Ocean Optics", v. 4341, 2000, p. 604-615.

[4] Бабиков Ю.Л и др., Интернет-коллекции по молекулярной спектроскопии, Сборник трудов 3 Всероссийской конференции по электронным библиотекам, Петрозаводск, 2001, с.183-187.

[5] Бабиков Ю.Л. и др., Сайт SPECTRA - информационный ресурс по молекулярной спектроскопии, VII Международная конференция по электронным публикациям "EL-Pub2002", http://www.ict.nsc.ru/ws/elpub2002/4384/

[6] D.W. Schwenke and H. Partridge, Convergence testing of the analytic representation of an ab initio dipol moment function for water: Improved fitting yields improved intensities, J.Chem.Phys., 113, No.16, 6592-6597, 2000.

[7] http://www.w3.org/TR/2003/WD-rdf-schema-20030123/

[8] http://www.w3.org/TR/2003/WD-owl-semantics-20030331/

[9] http://saga.atmos.iao.ru

[10] М.Р. Когаловский, Научные коллекции информационных ресурсов в электронных библиотеках, Труды первой Всероссийской конференции "Электронные библиотеки: перспективные методы и технологии, электронные коллекции", С.-Петербург, 1999, с. 16-31.

[11] Ахлёстин А.Ю., Гордов Е.П., и др., Интернет портал о свойствах атмосферы. Структура и технологии. Труды Всероссийской конференции "Математические и информационные технологии в энергетике, экономике и экологии", ч.2, Иркутск, 2003, с. 247-254.

Ontology for molecular spectroscopy of atmospheric gases

O.B. Rodimova, S.D. Tvorogov, A.Z. Fazliev

This work presents the description of the information resources on molecular spectroscopy, which are being developed at the Institute of Atmospheric Optics SB RAS. It gives the approach to the description of metadata for these resources by means of RDF for machine metadata processing as well as the ontology oriented for creation of a consistent interactive system and searching machines