Аннотирование информационных ресурсов в распределенной информационной системе "Молекулярная спектроскопия"

© А.В. Козодоев[1], А.И. Привезенцев[1,2], А.З. Фазлиев[1]

[1]Институт оптики атмосферы СО РАН, Томск
[2] Томский государственный университет систем управления и радиоэлектроники, Томск
faz@iao.ru

Аннотация

Аннотирование информационных ресурсов с помощью средств платформы XML является в настоящее время актуальной задачей. Формирование баз знаний, в основе которых лежат формализованные утверждения, позволяет обрабатывать содержание информационных ресурсов, расположенных на многочисленных сайтах, с помощью машины вывода. В нашей работе описан подход к формированию аннотаций для информационных ресурсов по молекулярной спектроскопии.

В работе дано краткое описание используемых структур данных и метаданных. Представлен подход к формированию онтологий для создания базы знаний в этой предметной области.

1 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 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 в сети Интернет с помощью ftp или пересылкой по почте на твердых носителях. Следствием сложившейся ситуации является прерывность поступления данных к пользователям, и практически полное отсутствие метаданных по причине неосведомленности спектроскопистов в возможностях современных информационных технологий.

В нашей работе описан подход к построению распределенной информационно-вычислительной системы (РИВС) в области молекулярной спектроскопии, основанный на использовании платформы XML для структурирования данных, описания и обработки метаданных, а также построения базы знаний и обмена знаниями. Практическое использование такой РИВС 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 в этой предметной области. Нижний Новгород, Москва, Санкт Петербург и Томск будут являться основными узлами этой системы. Заметим, что в настоящее время работает два автономных узла: http://saga.molsp.phys.spbu.ru, http://saga.atmos.iao.ru.

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 РИВС машинной обработке отдается значительное предпочтение. Это обстоятельство заставляет нас по иному отнестись к работе с метаданными и поставить задачу по разработке агентов и веб-

Труды 7[ой] Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL'2005, Ярославль, Россия, 2005.

сервисов. Создание почвы для таких разработок ния баз знаний в предметной области. Описание решения этой задачи является предметом данной работы.

Информационные системы, ориентированные на представления сервисов, характеризуют 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.

2 Уровень данных

Молекулярная спектроскопия является предметной областью относящейся к физике и для нее характерными являются два способа получения данных – эксперимент и расчет. Данные в молекулярной спектроскопии можно характеризовать тремя информационными структурами, которые связаны со структурными параметрами молекулы, параметрами спектральных линий и спектральными функциями [6]. Значительная часть физических величин, входящих в эти структуры, экспериментально измеряема. Стоит отметить, что в настоящее время наиболее изученной структурой данных являются параметры спектральных линий [7,8].

Данные и связанные с ними метаданные обычно называют информационными ресурсами. В этом разделе описаны операции над данными, которые используются при автоматическом составлении аннотации для источника данных.

Сложной задачей при работе с данными является составление банка данных по параметрам спектральных линий. Существующие банки параметров спектральных линий представляют из себя наборы строк, каждая из которых содержит значения более чем 10 физических величин. Заведомо не измеряемой характеристикой, входящей в строку, является идентификация спектральной линии, представляющая собой набор параметров, характеризующих принадлежность молекулы к классам (10 классов, связанных с квантовыми числами, определяющими колебательные состояния), группам (6 групп, связанных с квантовыми числами, описывающими вращения) и группам симметрии.

В реальных экспериментах измеряют, а в расчетах вычисляют, как правило, одну или две физические величины. Поэтому при составлении банка параметров спектральных линий с каждой строкой связывают несколько первичных источников данных. По такому правилу построены банки данных Hitran и Geisa.

требует построе
При коллективной работе с данными, когда пользователи заводят собственные первичные источники данных, важной становится процедура формирования составных источников данных, а значит и задача определения набора операций над источниками данных (фактически над данными, относящимися к этим источникам). Уникальным описателем строки в наборе параметров спектральных линий является идентификация линии. Другими словами, банки параметров спектральных линий не содержат разные версии значений, соответствующих набору, характеризующему идентификацию линии.

Набор типовых операций для данных в молекулярной спектроскопии ограничен и определён спецификой предметной области. В нашей работе мы ограничились рассмотрением логического представления данных и операции над однотипными данными в терминах реляционной модели данных [9]. Степень отношений, соответствующих определённому типу данных, будет одинаковой и равной количеству атрибутов, рассматриваемых для этого типа данных. Такие отношения будут совместимы между собой для бинарных операций.

Операции над данными можно разделить по количеству операндов на унарные и бинарные. Некоторые операции совпадают с классическим определением, а некоторые описаны с учётом специфики предметной области.

К числу унарных операций относятся выборка и проекция. Выборка – выбор из отношения кортежей, удовлетворяющих заданному условию. Новое отношение будет иметь ту же степень, а кардинальность будет меньше или равной кардинальности исходного отношения. Проекция – выбор из кортежей значений определённых атрибутов. При этом в зависимости от особенностей предметной области удаление дубликатов строк может не производиться. Новое отношение будет иметь такую же кардинальность, а степень меньше или равную степени исходного отношения.

Унарные операции применимы к источникам данных, содержащих строки произвольной структуры (в рамках понятий, связанных с параметрами спектральных линий).

К числу бинарных операций относятся объединение, коммутативное и некоммутативное соединение. При использовании бинарных операций главным ограничением является обязательное присутствие в строке параметров идентификации спектральной линии.

1. Объединение – конкатенация всех кортежей из исходных отношений в одно отношение с

удалением дублирующихся кортежей. Результирующее отношение будет иметь такую же степень, а кардинальность - равную или меньшую суммы кардинальностей исходных отношений. В случае с параметрами спектральных линий исходные отношения не должны пересекаться по спектральному диапазону и по колебательно-вращательной идентификации спектральных линий.

2. Соединение – это соединение по эквивалентности, выполненное по набору атрибутов. Степень отношения будет равна сумме степеней исходных отношений минус количество атрибутов, использующихся для соединения. В случае с параметрами спектральных линий соединение может производиться только по колебательно-вращательной идентификации спектральных линий. Возможен некоммутативный вариант соединения – левое или правое открытое соединение, когда из одного операнда выбираются все кортежи, а из другого, только те, что удовлетворяют условию соединения.

Предоставление пользователям интерфейса для манипуляций с данными, с одной стороны, дает пользователю возможность составления комплексных источников данных в рамках введенных операций, а с другой стороны, индуцирует необходимость включения в число метаданных операций над данными, которые характеризуют процесс формирования составного источника данных. Детальное знание истории формирования источника данных часто является одним из решающих факторов при сравнении результатов задач, полученных на их основе. Особо стоит отметить, что введенная нами процедура составления сложных источников данных, накладывает жесткое условие на обеспечение сохранности первичных источников данных.

3 Информационный уровень

Создание рекомендаций RDF и OWL в W3C заставило переосмыслить роль метаданных в информационных системах. Создатели рекомендаций построили базис для формирования инструментов и сервисов для разработчиков информационных систем с целью “проектирования и реализации высококачественных, значимых, корректных, минимально избыточных и хорошо аксиоматизированных онтологий” [9]. На базе этих онтологий должна быть решена следующая ключевая задача – создание машинно-обрабатываемых аннотаций для информационных ресурсов в вебе.

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 процессами 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 или OWL описание данных задачи, методов решения, результатов решения и т.д. Необходимыми компонентами к аннотациям являются онтологии задач, предметной области [11], качества и значений. Множество аннотаций, получаемых в результате ввода данных и решения задач, составляют базу знаний молекулярной спектроскопии. Здесь под базой знаний понимается множество утверждений на формальном языке (RDF) [12].

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

На рис.1 показаны аннотации, сгенерированные из базы знаний для чтения человеком. Существенное различие между ними состоит в наличии у аннотации к экспериментальным данным значений параметров регистрирующей аппаратуры и геометрии экспериментальной установки, тогда как у расчетных данных выделены параметры, характерные для используемого метода расчета, например, типа контура линии. Существенно разной является структура метаданных по источникам данных, относящихся к описанию информационного ресурса. Для экспериментальных данных – это ссылка на публикацию, тогда как для расчетных данных - это гиперссылка на комплексный источник данных.

Политика формирования базы знаний в ИВС основана на экспертном подходе. Пользователь, сформировавший информационный ресурс (на основе решенной задачи или ввода данных), устанавливает статус ресурса – направлен на публикацию. Эксперты принимают окончательное

решение о публикации. При положительном решении аннотации рекомендованного к публикации документа заносятся в базу знаний, а сам документ - в электронную библиотеку. Независимо от мнения экспертов пользователь может отправить ссылку на созданный им ресурс в коллекцию электронных документов, ассоциированную с ИВС в которой создан данный ресурс.

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	

a

Annotation DublinCore XML

Substance	
Absorbing gas	H2O
Broadening gas	H2O

Data array	
Wave number (number of values, unit) (cm^{-1})	4230
Absorption coefficient (number of values, unit) (cm^2/mol, calc)	4230
Errors	NO

Thermodynamical Conditions	
Temperature (°K)	575
Pressure (atm)	1
Broadening gas pressure (atm)	1

Spectral parameters	
Frequency range (cm^{-1})	2.050000e+03-2.249953e+03
Intensity (cm/mol)	> 1e-26

Data sources	
Data source of statistical sums	HITRAN 2001
Data source of spectral lines	HITRAN
Number of lines in spectral range	_
Number of lines is used under calculation	423

Approximations under absorption coefficient calculation	
Truncated contour (cm^{-1})	20
Contour type	Lorentz contour
Fragmentation method	Fragmentation factor

b

Рис.1. Визуализация утверждений, сформированных при вводе экспериментальной информации (a) и при расчете (b) коэффициента поглощения.

4 Уровень знаний

Данный уровень “действует как инфраструктура для поддержки управления и применения научного знания с целью достижения некоторых целей” [6]. Основной задачей для формирования связи между этим уровнем и двумя предыдущими является соотнесение масштабов контекста, с которым

проводится работа. Управление знанием тесно связано с жизненным циклом знаний, описание которого дано в [6].

Ключевым средством для описания знаний на этом уровне являются онтологии. Следуя работе [6], мы создали онтологию молекулярной спектроскопии (первая версия опубликована в работе [11]), онтологию задач молекулярной спектроскопии и онтологию значений. На начальном этапе исследований подготовка аннотаций осуществлялась в рамках рекомендации RDF [13]. В настоящее время для создания аннотаций информационных ресурсов применяется рекомендация OWL DL [14].

При построении онтологии по молекулярной спектроскопии мы исходили из следующей трехуровневой модели представления знаний предметной области. Базовым уровнем являются физические понятия, следующим уровнем являются математические модели физических понятий и третьим уровнем - программная реализация математических моделей в нашей ИВС. Степень детализации описания каждого из уровней определялась в рамках построенной онтологии персонализации (описание потребностей пользователя). В созданной нами онтологии по молекулярной спектроскопии мы исходили из того, что запросы пользователя ориентированы, прежде всего, на первый уровень, т.е. на усвоение знаний в рамках физических понятий. Два других уровня являются вспомогательными, раскрывают некоторые детали фактической реализации используемых математических моделей.

Онтология задач молекулярной спектроскопии включает в себя три базовых класса задач. К ним относятся задачи определения и измерения компонент энергии молекулы, задачи определения и измерения параметров спектральных линий и задачи определения и измерения спектральных функций. Эти задачи самым тесным образом связаны со структурами данных, описанными в начале этой статьи.

Общими задачами, решение которых можно ожидать в рамках созданной базы знаний, являются поиск и классификация ресурсов в РИВС, формирование системы поддержки решений и программного агента для работы с веб-сервисами. В настоящее время в ИВС закончены работы по организации семантического поиска в базе знаний. Некоторые детали решения этой задачи приведены ниже.

При написании онтологии ориентиром для нас была работа J.Sowa [15], позволившая с пониманием использовать категории верхнего уровня при формировании всех трех уровней (физический, математический и программный) представления знаний. Опишем для примера часть онтологии задач, связанную с задачей определения и измерения коэффициента поглощения. Ключевыми классами онтологии являются класс “Вещество” и около пятнадцати его подклассов, таких как “Уширяющий газ”, “Поглощающий газ”, классы значений таких свойств вещества, как “Температура”, “Давление”, “Частота перехода” и т.д. При формировании классов использованы все четыре аксиомы схемы. Например, класс, связанный со значениями такого свойства, как “Температура”, формировался как пересечение класса, образованного с помощью ограничения на свойство “иметьМатематическуюМодель”, с классом, образованным с помощью ограничения на свойство “иметьЕдиницуИзмеренияГрадусКельвина”.

Показанные на рис.1 аннотации информационных ресурсов являются в нашей онтологии индивидуалами, которые создаются как при вводе в систему экспериментальных данных, так и при проведении расчетов.

После выбора диалекта языка OWL DL нужно было определить: какой из существующих API использовать для создания своего программного обеспечения, работающего с онтологиями, какую применять машину вывода и какой пользовательский интерфейс машины вывода является наиболее удобным. Наша практика работы с онтологиями показала, что наиболее удобным пользовательским интерфейсом обладает машина вывода, встроенная в редактор онтологий Protégé 3.1. Но этот редактор работает с онтологиями OWL DL через plugin, а в этом OWL plugin нет встроенной машины вывода. Разработчики программы Protégé предлагают использовать для работы с OWL DL онтологиями машины вывода, реализующие DIG (DL Implementation Group) интерфейс. Отметим, что русскоязычный общий обзор инструментов инженерии онтологий можно найти, например, в [16].

Два основных открытых интерфейса прикладного программирования для работы с онтологиями API Jena [17] и OWL API [18] имеют практически одинаковую функциональность. В нашей работе [6] мы уже использовали интерфейс Jena для обработки RDF документов. Рассмотрим более подробно возможности API Jena по работе с онтологиями OWL DL. В частности, нас больше интересуют возможности подключения машин вывода, используя DIG интерфейс. В дистрибутиве Jena2 имеется несколько встроенных машин вывода:

- Transitive reasoner – Обеспечивает поддержку хранения и пересечения классов и свойств. Это реализация только свойств транзитивности и симметрии из rdfs:subPropertyOf и rdfs:subClassOf.

- RDFS rule reasoner – Реализация с конфигурируемым подмножеством RDFS импликации.
- OWL, OWL Mini, OWL Micro Reasoners – Набор полезной, но неполной реализации OWL/Lite подмножетсва языка OWL/Full.
- DAML micro reasoner – Используется внутри, чтобы дать возможность наследовать DAML API и обеспечивать RDFS расширение процесса логического вывода, схожий с имеющемся в Jena1.
- Generic rule reasoner – Норма, основывающаяся на машине вывода, чтобы поддерживать правила пользователя.
-

Как видно ни одна из имеющихся в Jena2 машин вывода не поддерживает полностью OWL DL. По этому поводу в документации по Jena сказано, что для полной поддержки вывода OWL DL нужно использовать внешние машины вывода, такие как Pellet [19], Racer [20] или FaCT [21]. Отметим, что использование Jena DIG интерфейса делает легким установку соединения с любой машиной вывода, поддерживающей DIG стандарт.

Из существующих машин вывода рассмотрим: FaCT, RacerPro, Pellet.

Машина вывода FaCT (Fast Classification of Terminologies) – это разработка профессора Манчестерского университета Яна Хоррокса (Ian Horrocks). FaCT – бесплатная DL машина вывода (DL reasoner), первоначально реализованная на Common Lisp, в настоящее время имеет несколько реализаций, в первую очередь - реализация DIG интерфейса для этой машины вывода, называемый FaCT DIG servlet [21].

Машина вывода RacerPro – это разработка немецкой компании Racer Systems GmbH & Co. KG. RacerPro – семантическое промежуточное программное обеспечение для индустриальных проектов, базирующееся на стандартах W3C RDF/OWL. Программа RacerPro имеет несколько видов лицензий; на бессрочное пользование продуктом выдаются только коммерческие лицензии.

Машина вывода Pellet – это разработка группы исследователей Semantic Web из лаборатории MIND LAB Мэрилендского университета. Pellet - open-source Java машина вывода OWL DL [19]. Эта машина вывода может совместно использоваться с API Jena и OWL API.

Подводя итог выше сказанному, выберем машину вывода OWL DL, исходя из наших потребностей. Машина вывода OWL DL должна быть бесплатной и иметь open-source коды на Java, реализовывать DIG интерфейс для совместимости с Jena.

Общим недостатком информационных систем по спектроскопии, развиваемых в ИОА СО РАН, является то, что они представляют систему знаний, сформированную одной научной школою. Создание распределенной информационной системы позволит выйти за рамки этого ограничения путем взаимообмена содержанием баз знаний. Технологией обмена утверждениями из БЗ в распределенной системе является репликация баз данных в которых они хранятся.

5 Благодарность

Авторы благодарны РФФИ за финансовую поддержку работы (грант № 05-07-90196).

Литература

[1] Tennyson et al., High accuracy *ab initio* rotation-vibration transitions of water, Science, **299**, 539-542, 2003.

[2] Бабиков Ю.Л., Барб А., Головко В.Ф., Тютерев Вл.Г., Интернет-коллекции по молекулярной спектроскопии, Сборник трудов 3 Всероссийской конференции по электронным библиотекам, Петрозаводск, 2001, с.183-187.

[3] S.Mikhailenko, Yu.Babikov, Vl.G.Tyuterev, A.Barbe, The DataBank of Ozone Spectroscopy on WEB (S&MPO), Computational Technologies, **v.7**, pp.64-70, (2002)

[4] Быков А.Д., Воронин Б.А., Козодоев А.В., Лаврентьев Н.А., Родимова О.Б., Фазлиев А.З., Информационная система по молекулярной спектроскопии. 1. Работа с данными, Оптика атм. и океана, 2004, **т. 17**, № 11, стр.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, 78p.

[6] Козодоев А.В., Привезенцев А.И.., Фазлиев А.З., Организация информационных ресурсов в распределенной информационно-вычислительной системе, ориентированной на решение задач молекулярной спектроскопии, Вычислит. Технологии, Специальный выпуск, 2005, **т**.10, с.82-91.

[7] HITRAN, http://cfa-www.harvard.edu/hitran/

[8] GEISA, http://www.ara.polytechnique.fr

[9] Конноли Т., Бегг К., Страчан А. Базы данных: проектирование, реализация и сопровождение. Теория и практика, 2-е изд.: пер. с англ.: М.: Издательский дом «Вильямс», 2000.-1120с.

[10] Ian Horrocks and Alan Rector, Foundations of the Semantic Web. Ontologies and OWL, http://www.cs.man.ac.uk/~horrocks/Teaching/cs646/Slides/introduction.pdf

[11] Родимова О.Б., Творогов С.Д., Фазлиев А.З., Онтология по молекулярной спектроскопии атмосферных газов, Труды 5 Всероссийской

конференции “Электронные библиотеки: перспективные методы и технологии, электронные коллекции”, С-Пб., 29-31 октября 2003, с. 211-215.
[12] Enrico Franconi, Description Logics, http://www.cs.man.ac.uk/~franconi
[13] Resource Description Framework (RDF): Concepts and Abstract Syntax, W3C Recommendation 10 February 2004, http://www.w3.org/TR/2004/REC-rdf-concepts-20040210/
[14] OWL Web Ontology Language Semantics and Abstract Syntax, W3C Recommendation 10 February 2004, http://www.w3.org/TR/2004/REC-owl-semantics-20040210/
[15] Sowa John F., *Knowledge Representation: Logical, Philosophical, and Computational Foundations*, Brooks Cole Publishing Co., Pacific Grove, CA, 2000. 594 p (Top-Level Categories, http://www.jfsowa.com/ontology/toplevel.htm)
[16] Овдей О. М., Проскудина Г. Ю., Обзор инструментов инженерии онтологий, http://www.impb.ru/~rcdl2004/cgi/get_paper_pdf.cgi?pid=26
[17] HP Labs Semantic Web Research, http://www.hpl.hp.com/semweb/
[18] OWL API, http://owl.man.ac.uk/api.shtml
[19] Pellet OWL Reasoner, http://www.mindswap.org/2003/pellet/index.shtml
[20] RacerPro version 1.8.1, http://www.racer-systems.com/
[21] The FaCT System, http://www.cs.man.ac.uk/~horrocks/FaCT/

Annotating the Information Resources in the Distributed Information System on Molecular Spectroscopy

A. Z. Fazliev[1], A. V. Kozodoev[1], A. I. Privezentsev[1,2]
[1]Institute of Atmospheric Optics, Tomsk, Russia
[2]Tomsk State University of Control Systems and Radio Electronics, Tomsk, Russia

Annotating the information resources in the frame of XML platform is a topical problem. The formation of knowledge base on the basis of formal statements allows one to process semantically the contents of information resources distributed in the Internet with help of the reasoners. In this paper the approach to description of molecular spectroscopy information resources is described.

The brief review of data structures and metadata used in molecular spectroscopy is presented. Ontology and knowledge base on molecular spectroscopy, is also discussed.