Логическая молекулярная спектроскопия

Привезенцев А.И.[1], Фазлиев А.З.[1], Tennyson J.[2]

[1]*Институт Оптики Атмосферы СО РАН, пр. Академический, д. 1, г. Томск, 634021, Россия.*
[2]*Department of Physics and Astronomy, University College London, Gower Street, London WC1E 6BT, United Kingdom*

remake@iao.ru, faz@iao.ru, j.tennyson@ucl.ac.uk

Аннотация. *Рассматривается база знаний, полученная в результате решения научной проблемы построения онтологических баз знаний для информационных систем, ориентированных на сбор достоверных первичных данных с целью систематизации знаний в предметной области и подготовки информации для решения прикладных задач в смежных предметных областях. Терминологическая часть (T-Box) базы знаний создана для связи с вычисляемыми метаданными и алгоритмами, которые позволяют автоматически формировать фактологическую часть (A-Box) базы знаний. Представлены возможности открытого представления машинно-обрабатываемых знаний и использование машины вывода.*

Ключевые слова: онтология, молекулярная спектроскопия, база знаний, информационная система

1 Введение

Доклад посвящен обобщению результатов работы разных исследовательских групп по обработке информационных ресурсов, относящихся к спектроскопии воды в рамках нескольких проектов. В Европе эти работы велись под эгидой международного союза чистой и атмосферной химии (IUPAC) [1] и в рамках гранта ИНТАС [2], в России - в рамках грантов РФФИ [3-5].

Молекулярная спектроскопия, являясь широко используемым разделов физики, характеризуется постоянно растущим огромным количеством спектральных данных. Работа с такими массивами данных требует, с одной стороны, предметной систематизации данных, а с другой стороны, программных средств для их обработки. Для работы с ними создаются специальные базы постоянно пополняющихся спектральных данных: HITRAN, GEISHA, VALD, CDMS, BASECOL, STSP. На основе этих баз данных создаются информационные системы. Для молекулярной спектроскопии характерны следующие информационные проблемы: базы данных могут содержать недостоверные данные; имеется неполнота информации о собранных данных, об их способах получения; существующие информационные системы не дают средств для автоматизированного программного анализа этой информации. Для решения этих проблем и задач классификации, интеграции, поиска и сравнения информационных ресурсов была создана база знаний в рамках научной информационно-вычислительной системы. Отчет по выполненной работе, относящейся к задачам спектроскопии воды, будет публиковаться в печати в течение нескольких следующих лет, а первая статья этого цикла публикаций появится в 2009 [6].

2 Информационные аспекты молекулярной спектроскопии

Информационные аспекты спектроскопии воды, следующие из информационной модели, изложены в докладе [7] и опубликованы в монографии [8]. Информационная модель молекулярной спектроскопии необходима для создания терминологической части базы знаний. В зависимости от информационных задач решаемых с помощью базы знаний необходимо использовать адекватную им информационную модель. Так как большая часть информационных задач связана со знаниями, основанными на решениях задач предметной

области [9] используем информационную модель молекулярной спектроскопии в виде цепей её прямых и обратных задач. Основной акцент для модели предметной области в виде цепей прямых и обратных задач сделан на автоматизируемое установление достоверности данных решений задач. Описания результатов решений задач предметной области при представлении знаний рассматриваются как наборы фактов (фактологическая компонента). Задачи классификации в молекулярной спектроскопии, как правило, сводятся к построению таксономии терминов предметной области и на практике рассматриваются как вспомогательные задачи. Предполагается, что в задаче классификации концепты представляют понятия предметной области (терминологическая компонента). Систематизация предметной области проведена на нескольких терминологических таксономиях и наборах фактов для каждой из задач модели предметной области. Кроме этого, важную роль при создании базы знаний играет концепт “информационный источник”, который содержит данные, опубликованные в одной статье и относящиеся к молекуле одного типа и необходим при семантическом описании решений задач молекулярной спектроскопии.

Прямые задачи молекулярной спектроскопии связаны с расчетами из первых принципов фундаментальных характеристик молекул, таких как уровни энергии молекул, частоты перехода, коэффициенты Эйнштейна и т.д. Обратные задачи молекулярной спектроскопии связаны с обработкой данных измерений спектральных функций, что позволяет в дальнейшем при машинной обработке классифицировать их выходные данные как экспериментальные. В цепи задач молекулярной спектроскопии существуют связи между прямыми и обратными задачами.

При решении задач обоих типов проводятся вычисления одних и тех же физических величин. Их сравнение между собой позволяет делать выводы о достоверности данных.

К классам элементарных прямых задач, используемых нами для проектирования информационной системы, относятся следующие классы задач: задача определения физических характеристик изолированной молекулы (T1); задача определения параметров спектральной линии изолированной молекулы (T2); задача определения параметров контура спектральной линии (T3); задача расчета спектральных функций (T4).

Данная цепь задач определяет последовательность их решения. Например, входные данные задачи T3 должны включать в себя выходные данные задачи T2. Выделение первых двух классов задач (T1 и T2) обусловлено важным физическим фактором, а именно, свойства изолированных молекул не зависят от термодинамических параметров. Задача T3 позволяет определить параметры спектральных линий одной молекулы при реперных термодинамических условиях (температуры 296K, 1000K, 3000K) и учете столкновений молекул в газе. Задача T4 описывает излучательные или поглощательные способности газов при разных термодинамических условиях.

К классам элементарных обратных задач, используемых нами для проектирования информационной системы, относятся следующие классы: задача измерения спектральных функций (E1); задача приписывания квантовых чисел спектральным линиям (T5); задача определения коэффициентов Эйнштейна (T6); задача определения уровней энергии изолированной молекулы (T7).

В информационной системе возможно проведение сравнительного анализа решений однотипных задач по одинаковым физическим величинам. Так, результаты решения для задачи T1 сравнимы с результатами задачи T7, по уровням энергии. Решения T2 сравнимы с результатами решений T6, по переходам. Решения задача T3 сравнимы с результатами решений T5, по интенсивностям, полуширинам, сдвигам и термодинамические зависимостям полуширин и сдвигов. Решения задача T4 сравнимы с E1, по значениям спектральных функций.

3 Прикладная онтология молекулярной спектроскопии

В контексте данной работы понятие «онтология» и «база знаний» равнозначны и используются в соответствии с определением языка спецификации онтологий OWL DL. Язык OWL DL имеет формальный синтаксис и формальную семантику дескриптивной логики SHOIN(D). Для машинной обработки онтологий существует несколько машин вывода, основанных на дескриптивной логике, например FaCT++, Pellet.

Созданные прикладные онтологии по молекулярной спектроскпии используются для машинной систематизации и интерпретации знаний, для интеграции знаний в другие смежные предметные области, а также организации семантического поиска. Использование онтологии возможно с помощью любого редактора онтологий, например использование Protégé показано на рис. 1.

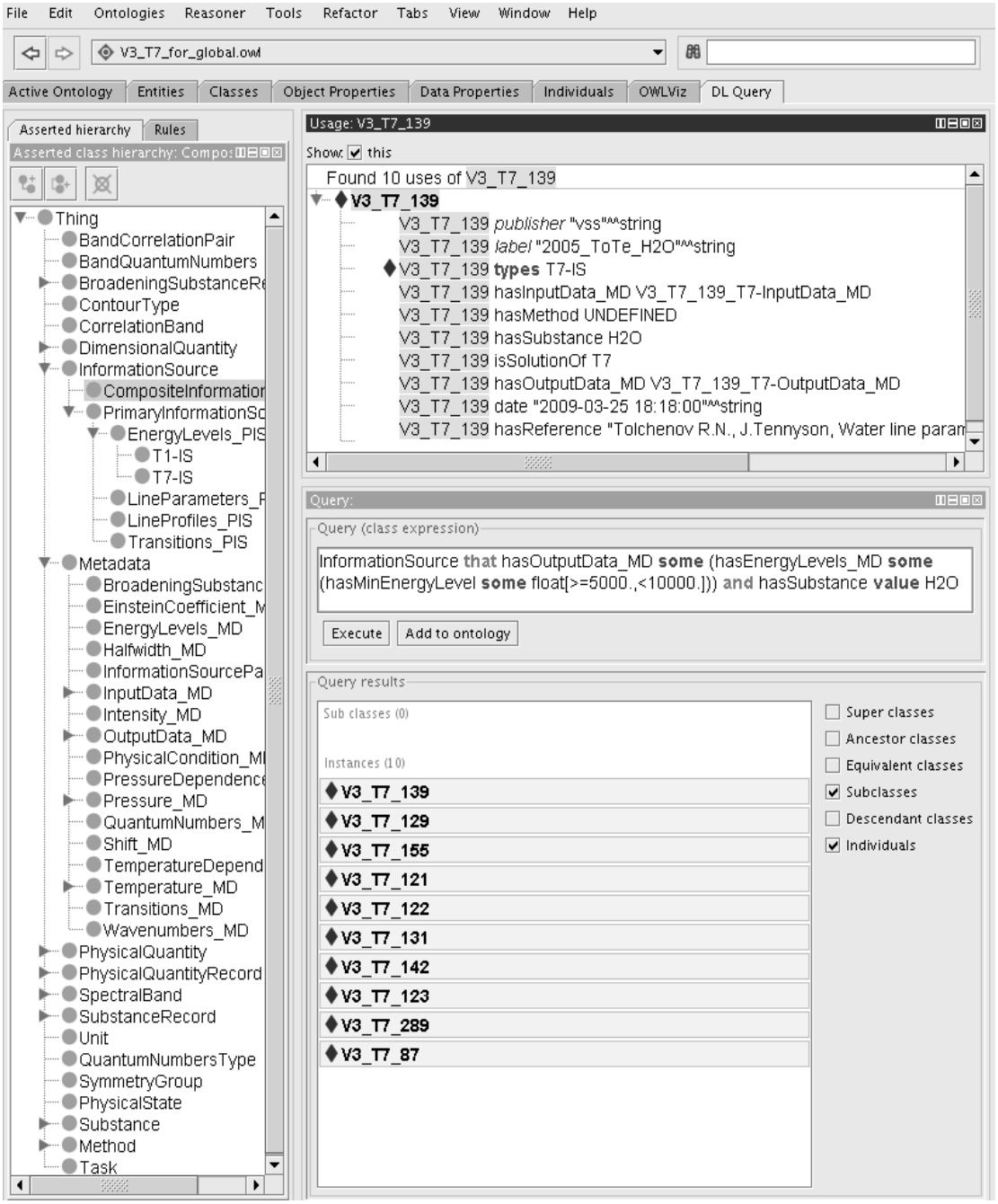

Рис.1. Программный интерфейс Protégé для работы с базой знаний.

В левой части рис. 1 показана таксономия терминологической компоненты онтологии по молекулярной спектроскопии, в которой представлены понятия(классы) информационной модели предметной области. Большая часть классов прикладной онтологии задач построены с помощью ограничений на свойства. Например, класс InformationSource содержит подклассы CompositeInformationSource, PrimaryInformationSource. Экземплярами класса PrimaryInformationSource являются наборы фактов о физических величинах относящиеся к одной молекуле и одной публикации. Составные источники могут содержать данные о нескольких молекулах, которые опубликованы в нескольких статьях. Основным свойством

онтологической базы знаний на языке OWL является открытость, то есть возможность любому человеку дополнять и изменять базу знаний под свои информационные задачи.

Семантический поиск по базе знаний можно осуществлять с помощью манчестерского синтаксиса языка OWL. На рис. 1 справа в центре показан пример поискового запроса: найти все источники информации по основному изотопомеру воды, где минимальный уровень энергии находится в интервале между 5000 и 10000. На рис. 1 справа внизу показаны экземпляры класса InformationSource, найденные в результате семантического поиска.

Особенностью работы с онтологиями задач в информационной системе «W@DIS»[7] является тот факт, что пользователи при решении задач механически составляют свою собственную онтологию, содержащую экземпляры классов, соответствующие конкретной решенной задаче. Эти индивиды можно объединять с онтологиями других задач или других пользователей, если это позволяют права доступа к базе знаний, например:

http://wadis.saga.iao.ru/saga2/meta/get/V3_T1+T7_for_global.owl&V3_T1_for_user.owl

Прикладная онтология по молекулам воды и углекислого газа содержит описание 881 первичного источника информации, содержащих решения 6 задач спектроскопии воды и углекислого газа и парных корреляций решений однотипных задач в виде утверждений, описанных средствами языка *OWL DL*. Статистические данные о числе высказываний относящихся к A-Box и T-Box приведены в таблице 1.

Таблица 1. Количественные характеристики базы знаний молекулярной спектроскопии.

Задача	**Источники информации H_2O**	**Источники информации CO_2**	**Триады (A-box)**	**Триады корреляций (A-box)**	**Триады (T-Box)**
T1	26	0	1082	110740	
T7	156	0	5568		
T2	30	23	2937	467205	
T6	351	6	16033		
T3	27	7	2510	280742	
T5	254	1	19537		
Всего	844	37	47667	858687	1943
Итого	881		908297		

Исследованная прикладная онтология источников информации о свойствах решения задач молекулярной спектроскопии, индуцировала решение следующей задачи – задачи построения логической теории молекулярной спектроскопии. Первые результаты выполненного нами исследования представлены в докладе. Выполненная работа использует методы представленные Зиновьевым А.А. [10, 11].

4 Заключение

Основным преимуществом использования онтологической базы знаний на языке OWL DL по описанию результатов решений задач в молекулярной спектроскопии, является возможность использования универсального формата обмена знаниями в Web с явной спецификацией знаний в предметной области. Кроме этого на онтологической базе знаний возможна машинная логическая проверка с помощью машины вывода на достоверность и неполноту информации о собранных данных в виде фактов описания результатов решений задач, следующих категорий: проверка ограничений по заданным правилам, следующие из математических моделей

исследуемых физических объектов (квантовые числа); систематизация решений задач по величине среднеквадратичные отклонения; систематизация информации о частях составных источников данных; указание недостоверных данных, выявленных экспертом предметной области.

Литература

[1] Complete Spectroscopy of Water: Experiment and Theory**",** INTAS grant 03-51-3394

[2] IUPAC project No.2004-035-1-100 “A database of water transitions from experiment and theory”
http://www.iupac.org/web/ins/2004-035-1-100

[3] SPECTRA - информационно-вычислительная система по молекулярной спектроскопии, грант РФФИ 02-07-90139

[4] Распределенная ИС "Молекулярная спектроскопия" грант РФФИ № 05-07-90196

[5] Интернет доступная информационная система по молекулярной спектроскопии, основанная на знаниях, грант РФФИ №08-07-00318-а

[6] J. Tennyson, P.F. Bernath, L.R. Brown, A. Campargue, M.R. Carleer, A.G. Császár, R.R. Gamache, J.T. Hodges, A. Jenouvrier, O.V. Naumenko, O.L. Polyansky, L.S. Rothman, R.A. Toth, A.C. Vandaele, N. Zobov, L. Daumont, A.Z. Fazliev, T. Furtenbacher, I.F. Gordon, S.N. Mikhailenko, S.V. Shirin, IUPAC Critical Evaluation of the Rotational-Vibrational Spectra of Water Vapor. Part I. Energy Levels and Transition Wavenumbers for $H_2{}^{17}O$ and $H_2{}^{18}O$ // J.Quant.Spectr.Rad.Transfer, 2009, DOI: 10.1016/j.jqsrt.2009.02.014.

[7] A.Z.Fazliev, A.G.Császár, J.Tennyson, W@DIS: Water spectroscopy with a Distributed Information System // Proc. of the 10 HITRAN Database Conference, 2008, p.38-39

[8] Быков А.Д., Науменко О.В., Родимова О.Б., Творогов С.Д., Тонков М.В., Фазлиев А.З., Филиппов Н.Н., Информационные аспекты молекулярной спектроскопии, Из-во ИОА СО РАН, 2008, 356с.

[9] A.D.Bykov, A.Z. Fazliev, N.N.Filippov, A.V. Kozodoev, A.I.Privezentsev, L.N.Sinitsa, M.V.Tonkov and M.Yu.Tretyakov, Distributed information system on atmospheric spectroscopy // Geophysical Research Abstracts, SRef-ID: 1607-7962/gra/EGU2007-A-01906, 2007, v. 9, p. 01906.

[10] Зиновьев А.А., Основы логической теории знаний, М., Наука, 1967, 260с.

[11] Зиновьев А.А., Логическая физика, М., Наука, 1972, 191с.