Инструментарий публикации данных и метаданных для распределенной информационной системы по количественной спектроскопии♣

© А.Ю. Ахлёстин, Н.А. Лаврентьев, М.М. Макогон, А.И. Привезенцев, А.З. Фазлиев

Институт оптики атмосферы СО РАН имени В.Е. Зуева, г. Томск
faz@iao.ru

Аннотация

Описаны некоторые компоненты инструментария публикации данных и метаданных, созданного в проекте «Виртуальный центр атомных и молекулярных данных» (VAMDC). Две из них, связанные с проверкой достоверности данных и манипуляцией с данными, описаны детально.

Рассмотрены проверки двух типов ограничений. К первому из них относятся ограничения на значения физических величин, следующие из математических моделей процессов и объектов количественной спектроскопии, в частности, правил отбора. В качестве примера рассмотрены результаты проверки данных из более 700 статей по спектроскопии воды.

Другой тип ограничений связан с фактом публикации данных (ограничения существования). Эти ограничения применялись для части массива данных HITRAN, относящихся к изотопомерам молекулы воды. Декомпозиция, используемая для проверки этого ограничения, основана на использовании полного набора опубликованных данных, собранного группой данных проекта IUPAC (Международный союз чистой и прикладной химии).

1 Введение

Научная статья возникает в результате синтеза ее авторами результатов исследований объектов и процессов предметной области. Обязательной ее компонентой являются новые утверждения об исследованных объектах, основанные на существующих данных и информации о предметной области.

Появление глобальной информационной системы (Web) значительно облегчило поиск и доставку данных и информации исследователям и инициировало работы по созданию виртуальных центров данных. В таких центрах (например, [1 – 7]) собираются, систематизируются, хранятся, публикуются и предоставляются пользователям решения задач предметных областей. Структура центров данных может быть различной: одни состоят из одной организации [1 – 6], другие из нескольких организаций, и в этом случае центр данных является распределенной информационной системой [7].

VAMDC является одним из таких центров, создаваемых в рамках европейского инфраструктурного проекта седьмой рамочной программы. Его основой является распределенная информационная система, содержащая распределенные неоднородные по интенсионалам данные по атомной и молекулярной спектроскопии. Каждый узел этой системы поддерживается автономно разными организациями. Виртуальность VAMDC подразумевает централизацию метаданных, связанных с данными распределенной информационной системы. Одной из частей метаданных является реестр данных распределенной системы [8], содержащий перечень информационных ресурсов, названия организаций, создавших и поддерживающих ресурсы и т. д. Наряду с этими традиционными метаданными, связанными с описанием абстрактного ресурса, существуют метаданные, характеризующие свойства данных, относящиеся к предметной области.

Для создания виртуального центра необходимо решить ряд прикладных задач, в частности, задачу построения инструментария публикации данных и метаданных, которая разбивается на ряд подзадач, некоторые из которых описаны в данной работе.

К числу этих подзадач относятся задача *вычисления* достоверности информационных ресурсов, публикуемых центром данных, и задача автоматизации манипуляций с данными на разных этапах их публикации и при формировании пользователем нужных ему структур данных.

Авторы являются участниками проекта VAMDC, в котором они разрабатывают систему публикации данных, метаданных для однородной по программному обеспечению распределенной информационной системы по количественной молекулярной спектроскопии, развиваемой ими в России [9].

2 Особенности предметной области

Целью VAMDC является построение «безопасной, документированной, гибкой, легко доступной и интероперабельной цифровой инфраструктуры для атомарных и молекулярных данных». После принятия участниками проекта XML-схемы [10], характе-

Труды 12[й] Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2010, Казань, Россия, 2010

ризующей структуру данных спектроскопии, количественный аспект спектроскопии оказался главным.

Количественная спектроскопия является сформировавшейся дисциплиной, относящейся к оптике. В ней сформировалась концептуализация и построены основы логической теории. Появляющиеся новые методы расчета значений спектральных характеристик используют новые базисные функции, что в свою очередь приводит к новой интерпретации результатов и дополнениям в концептуализацию. Такая ситуация является типичной для физических наук.

Особенностью количественной спектроскопии является огромное число данных, получаемых в расчетах значений физических величин (например, для изотопомера молекулы воды HDO [11] число переходов составляет около семисот миллионов). Характерной чертой получаемых в измерениях спектральных данных, является то, что в разных диапазонах частот переходов молекул и атомов используются устройства, основанные на разных физических процессах. Не существует устройства, позволяющего измерить с высоким разрешением спектральные функции молекулы во всем диапазоне изменений частоты излучения или поглощения. Это означает, что значения спектральных характеристик в разных диапазонах их изменений определяются с разной точностью.

Значительный объем данных рассматриваемой предметной области определяет структуру публикуемых статей, которая, как правило, включает таблицы, содержащие значения физических величин. Представление статей в цифровом виде привело к тому, что в последнее десятилетие получила широкое распространение практика публикации данных в виде файлов в приложении к статье. Наряду со стандартным механизмом публикации данных в текстах статей или приложений к ним используется размещение данных, которые из-за значительных размеров не публикуются издательствами, в интернет доступных информационных системах или ftp-серверах [12] организаций, в которых проводятся исследования. Эти системы и являются составными частями распределенной информационной системы по количественной спектроскопии.

3 Публикация данных и инструментарий

Обратимся к анализу процесса публикации данных в таких системах и связанных с ним составных частей инструментария публикации данных. В информационной системе в цикле жизни данных и метаданных [13] этап публикации тесно связан с этапами приобретения данных системой и их извлечением из системы пользователями. По этой причине существующий механизм публикации статьи в цифровых журналах содержит

- систему ее загрузки и приложений к ней в информационную систему издательства;
- рецензирование статьи (в настоящее время рецензенты своими силами организуют проверку целостности данных и логических выводов автора);
- переписку рецензентов и редактора с авторами,
- механизм принятия решения о публикации и размещение статьи и приложений в информационной системе издательства или посредников;
- систему представления статей и приложений к ним.

Статья представляет собой информационный ресурс, который можно представить в виде трех частей: данных предметной области, метаданных и логической теории (онтологии), связанной с предметной областью этими данными и метаданными. Автоматизация работы с каждой из этих частей основана на использовании языков их спецификаций. В нашей работе такими языками являются XML, RDF и OWL DL [14].

Публикация данных в виртуальном центре данных позволяет сделать степень автоматизации процесса публикации существенно большей по сравнению с традиционной публикацией статьи. Во-первых, авторы могут до начала процедуры публикации предварительно загрузить собственные данные и ознакомиться с наборами свойств этих данных, созданными системой. Набор таких свойств включает в себя результаты проверок формальных ограничений, в том числе стандартные отклонения со всеми опубликованными данными (в настоящее время такие наборы данных в ИС W@DIS существуют для молекул воды, углекислого газа и сероводорода) и ограничения существования данных.

Автоматизация механизма публикации может расширить возможности следующих этапов, связанных с публикацией данных:

1. Загрузка структурированных данных и генерация предметных метаданных;
2. Рецензирование данных и автоматическое вычисление стандартных отклонений с данными, опубликованными за всю историю существования спектроскопии;
3. Извлечение данных и метаданных предметной области, в том числе детальное описание стандартных отклонений.

Расширение возможностей процедуры загрузки данных состоит в том, что к числу типовых метаданных (авторы, дата поступления в редакцию и т. д.), предназначенных для формирования реестра публикаций, автоматически добавляются:

- метаданные, описывающие структуру данных;
- метаданные, характеризующие интервалы изменения физических величин;
- метаданные, связанные с проверками ограничений (проверка целостности данных), следующими из математических моделей молекул;
- метаданные, определяющие парные отношения загруженных данных со всеми данными, имеющимися в системе;

- метаданные, описывающие мереологию [15] данных (не опубликованные ранее части данных, опубликованная часть данных и их принадлежность к публикациям и т. д.).

Расширение возможностей рецензирования заключается в упрощении для эксперта рутинной части проверок достоверности данных. За экспертом остается необходимость проведения неформальных проверок данных, связанных с ними метаданных и неформальных логических конструкций.

Наконец, расширение возможностей в извлечении данных пользователем состоит в том, что с реализацией сервисов, обеспечивающих ему манипуляции с данными, пользователю становятся доступными более сложные действия по формированию его личных массивов данных. Эти действия реализуются с помощью унарных и бинарных операций над пространственно распределенными данными. Система строится с помощью вебсервисов и основана на реестре информационных ресурсов, общем для всех узлов РИС.

4 Публикация данных и инструментарий

4.1 Интерпретация достоверности

Прежде всего, стоит отметить, что в разных предметных областях определение достоверности данных разное [16]. По этой причине программное обеспечение, реализующее проверку достоверности, будет разным для разных предметных областей. Определим, в каком смысле используется понятие достоверности информационных ресурсов предметной области в данной работе. В данной работе мы рассматриваем только такие определения достоверности данных, которые позволяют проводить компьютерную проверку достоверности.

Проверка достоверности связана с проверкой ограничений на информационные ресурсы предметной области. В работе рассмотрены две группы ограничений: ограничения на значения физических величин и ограничения на существование этих значений.

Первая группа ограничений связана с математическими моделями молекул и физическими ограничениями на рассматриваемые в предметной области процессы. Характерным для количественной спектроскопии примером являются правила отбора для переходов, следующие из математической модели молекулы. С формальной точки зрения этим ограничениям соответствует проверка истинности утверждения $(\forall X)D$, где X – данные, а D – предметная область.

Вторая группа ограничений связана с интерпретацией существования информационных ресурсов [17], которыми в количественной спектроскопии являются решения ее задач. Эти ограничения обусловлены тем фактом, что решения задач являются входными данными для приложений интернет доступных информационных систем, и если они не опубликованы (т. е. не имеют URI), то приложение не может их использовать. Другими словами, эти ресурсы для приложений не существуют. К числу несуществующих ресурсов в такой трактовке относятся неопубликованные решения, в том числе потенциально вычислимые по известным алгоритмам. Этой группе ограничений соответствует проверка истинности утверждения $(\exists X)D$.

4.2 Ограничения на значения

В количественной спектроскопии выбор исследователем математической модели молекулы означает выбор предметной области, а значит и ряда критериев достоверности данных. Заметим, что в количественной спектроскопии используются разные математические модели одной и той же молекулы, а значит и для проверки достоверности используются разные наборы критериев. Например, проверка допустимых интервалов изменения физических величин (вакуумных частот, интенсивностей, уровней энергии и т. д.), типов данных значений спектральных величин и соответствия квантовых чисел правилам отбора трактуется как проверка достоверности ограничений на значения.

Не для всех ограничений на значения можно построить разрешимый алгоритм проверки, т. е. такой, который выполняется компьютером за конечный интервал времени. Связано это с тем обстоятельством, что некоторые ограничения имеют эвристический характер и не являются формализуемыми. Принятие решения о соответствии данных этим критериям осуществляется экспертами предметной области.

Следовательно, инструментарий публикации, в части проверки достоверности на ограничения физических величин, должен содержать два набора программ. Один – для вычислений достоверности по формальным критериям, а другой – для ввода результатов экспертной оценки. Соответствующее программное обеспечение было создано для ИС W@DIS, представляющей информационные ресурсы, относящиеся к спектроскопии некоторых молекул, и использовалось для анализа достоверности данных из ~ 1400 статей о спектральных свойствах молекул воды, сероводорода и углекислого газа и их изотопомеров по критерию ограничения на значения.

Детали формирования модели предметной области, положенные в основу формирования программного обеспечения, были ранее описаны в [18, 19], а предварительные результаты работы представлены в работе [20].

Данные, отнесенные к одной молекуле и полученные одним методом, назовем источником данных. Источник данных, содержащий только данные, удовлетворяющие ограничениям на значения, будем называть приведенным к канонической форме.

В табл. 1 приведены результаты классификации источников данных по спектроскопии воды для двух групп прямых и обратных задач. В ней используются следующие обозначения. Первое число в колонке соответствует общему числу источников данных, а в

скобках – числу источников, содержащих только достоверные данные.

Как следует из табл. 1, только 60% публикаций, содержащих решения задач Т2, Т6 (определение частот перехода изолированной молекулы), не содержат ошибок, связанных с правилами отбора. Для решений задач Т3 и Т5 (определение параметров спектральных линий при нормальных условиях) процент публикаций, не содержащих ошибок, составляет 78%.

Таблица 1. Результаты проверки достоверности первичных источников информации о решениях задач спектроскопии воды [8]

Молекулы	Задача Т2, Т6	Задачи Т3, Т5
H_2O	5(0), 91 (47)	5 (0), 183 (167)
$H_2{}^{17}O$	5(1), 40 (31)	4 (0), 19 (16)
$H_2{}^{18}O$	5(1), 59 (35)	4 (0), 29 (17)
HDO	3(0), 83 (56)	2 (0), 8 (3)
$HD^{17}O$	2(0), 3 (3)	2 (0), 6 (6)
$HD^{18}O$	2(0), 6 (6)	2 (0), 7 (7)
D_2O	3(0), 38 (26)	3 (0), 10 (7)
$D_2{}^{17}O$	1(0), 3 (3)	2 (0), 1 (1)
$D_2{}^{18}O$	2(0), 6 (6)	2 (0), 1 (1)
	28(2), 318 (207)	26(0), 264 (225)

Стоит отметить, что анализ составных источников данных по спектроскопии, таких, как HITRAN [21] и GEISA, для ряда изотопомеров воды выявил наличие десятков переходов, не удовлетворяющих правилам отбора.

4.3 Ограничения существования

Решения задач спектроскопии, полученные исследователями, как правило, публикуются в печати или, в последние 10 лет, в интернете. На практике в научном сообществе спектроскопистов принято ссылаться на публикации только в ограниченном списке журналов или сайтов.

Назовем источник данных по количественной спектроскопии первичным источником данных, если все данные из него опубликованы в одной статье.

Предположим, что существует полный набор первичных источников данных в предметной области, и все источники данных в нем приведены к канонической форме. Для изотопомеров молекулы воды $H_2{}^{17}O$, $H_2{}^{18}O$, HDO, $HD^{17}O$ и $HD^{18}O$ такой полный набор описан в статье [22, 23].

Тогда уместна постановка ~~следующей~~ задачи о разложении (декомпозиции) произвольного массива спектральных данных по данным из полного набора первичных источников данных, приведенных к каноническому виду.

В спектроскопии сравниваемые физические величины должны характеризоваться одинаковыми квантовыми числами. Значения же сравниваемых физических величин могут отличаться, тем не менее, физическая сущность, описываемая этими величинами (например, переход), интерпретируется как одна и та же. Уточним критерий, по которому значения физических величин, относящиеся к одинаковому набору квантовых чисел и описывающие переход между состояниями, совпадают. Пусть ω_1 и ω_2 – сравниваемые частоты, характеризуемые одинаковым набором квантовых чисел. Будем считать их совпадающими, если их разность удовлетворяет неравенству

$$\left|\omega_1 - \omega_2\right| < \varepsilon . \qquad (1)$$

В количественной спектроскопии ε связана с разрешающей способностью измерительных устройств. Величина ε является разной для разных диапазонов значений сравниваемых частот.

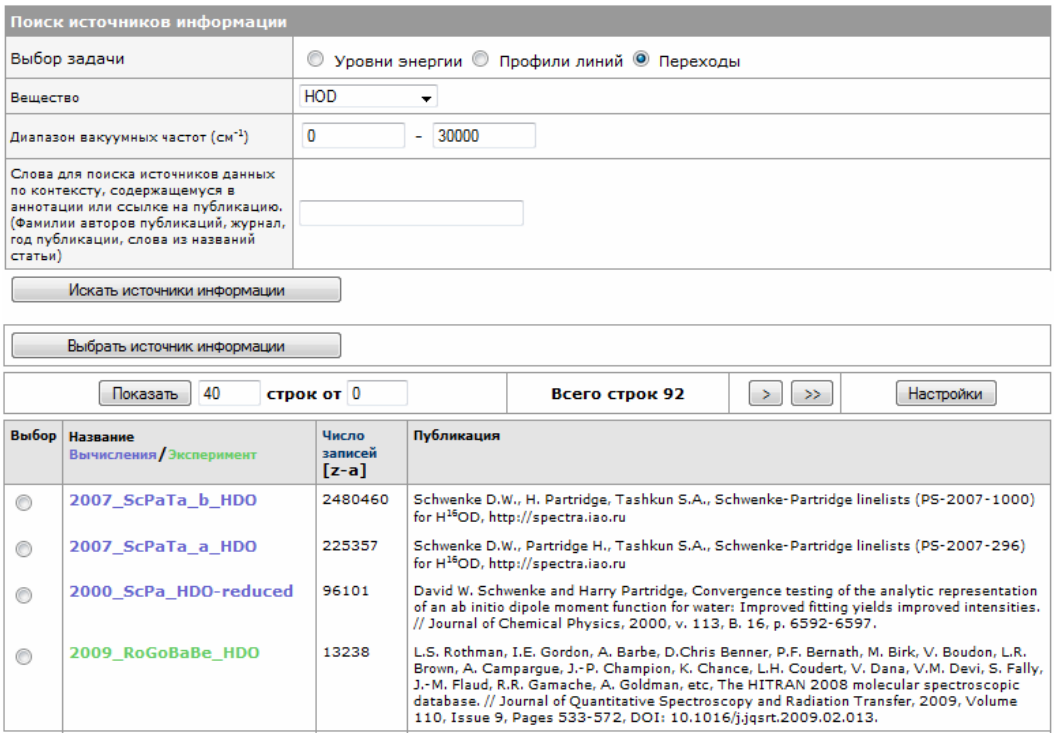
В качестве примера разложения экспертных массивов данных приведем фрагмент разложения массива спектральных данных для изотопомера воды HDO, взятого из банка данных HITRAN [21], выполненного с точностью ε в диапазоне частот перехода 0-20000 cm^{-1}. Интерфейс для выбора источников данных показан на рис. 1.

Декомпозиция проведена для разных интервалов, т. к. точность измерения в них является разной. В каждом из интервалов декомпозиция проводилась отдельно по источникам данных, содержащих данные измерений (верхняя строка подраздела таблицы) и вычислений (нижняя строка).

На рис. 2 показан интерфейс для задания допустимой ошибки определения частоты (вакуумных волновых чисел) ε и выбора способа разложения (разложение по данным расчета и/или данных измерений).

При разложении массива HITRAN по данным измерений молекулы HDO получено, что число не содержащихся в публикациях переходов равно 2212, из общего числа 13238, что составляет 17%. Разложение по опубликованным данным измерений, дает остаток, равный 3704 переходам. Разложение по публикациям, содержащим данные вычислений из первых принципов, дает остаток в 63 перехода. Заметим, что точность в десятую долю обратного сантиметра является очень грубой во всем спектральном диапазоне измеряемых частот перехода, тогда как для вычисляемых из теории значений она является удовлетворительной.

Проверка ограничения существования информационных ресурсов при создании инструментария публикации в центре данных, необходима в первую очередь для экспертов, принимающих решения о возможности публикации того или иного массива спектральных данных. Она также может быть полезна при планировании экспериментов для уточнения значений физических величин, описывающих спектры молекул.



Поиск источников информации

Выбор задачи	Уровни энергии / Профили линий / Переходы
Вещество	HOD
Диапазон вакуумных частот ($см^{-1}$)	0 - 30000
Слова для поиска источников данных по контексту, содержащемуся в аннотации или ссылке на публикацию. (Фамилии авторов публикаций, журнал, год публикации, слова из названий статьи)	

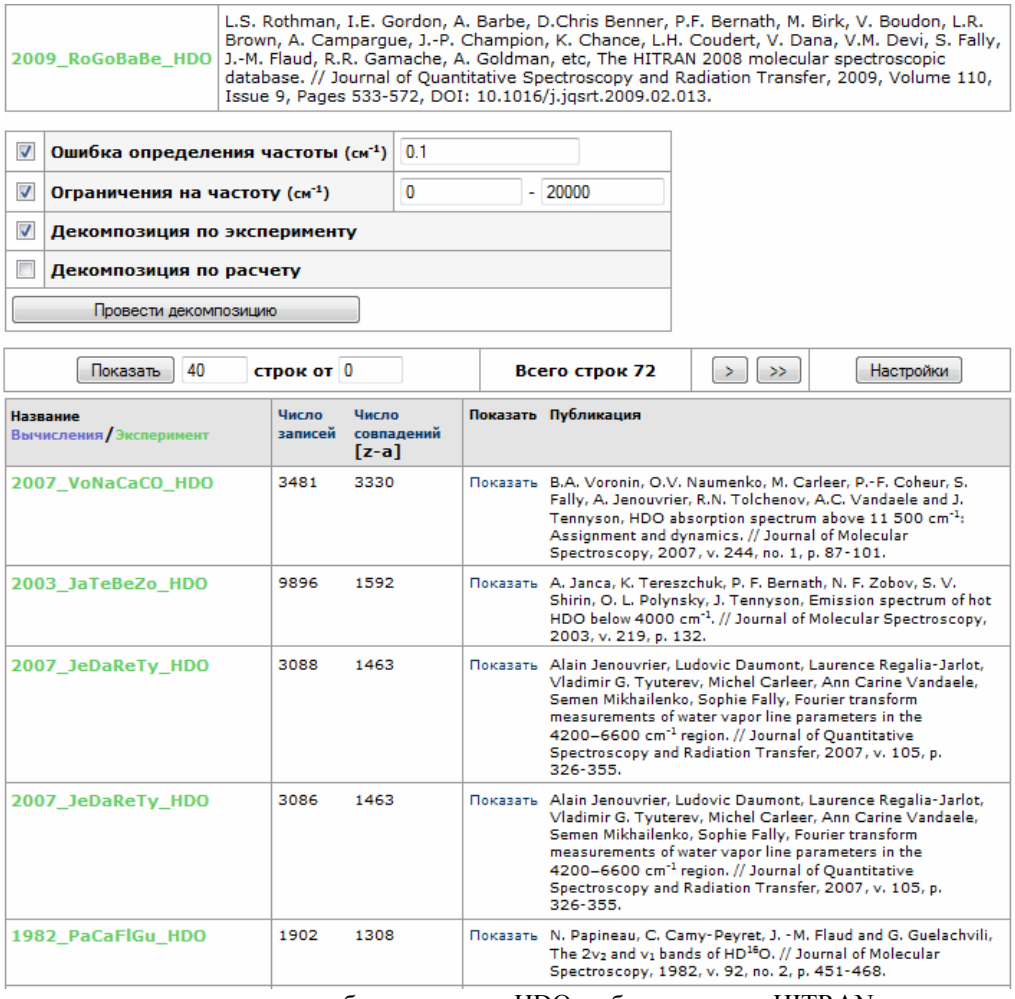
Искать источники информации

Выбрать источник информации

Показать 40 строк от 0 | Всего строк 92 | > >> | Настройки

Выбор	Название Вычисления/Эксперимент	Число записей [z-a]	Публикация
○	2007_ScPaTa_b_HDO	2480460	Schwenke D.W., H. Partridge, Tashkun S.A., Schwenke-Partridge linelists (PS-2007-1000) for $H^{16}OD$, http://spectra.iao.ru
○	2007_ScPaTa_a_HDO	225357	Schwenke D.W., Partridge H., Tashkun S.A., Schwenke-Partridge linelists (PS-2007-296) for $H^{16}OD$, http://spectra.iao.ru
○	2000_ScPa_HDO-reduced	96101	David W. Schwenke and Harry Partridge, Convergence testing of the analytic representation of an ab initio dipole moment function for water: Improved fitting yields improved intensities. // Journal of Chemical Physics, 2000, v. 113, B. 16, p. 6592-6597.
○	2009_RoGoBaBe_HDO	13238	L.S. Rothman, I.E. Gordon, A. Barbe, D.Chris Benner, P.F. Bernath, M. Birk, V. Boudon, L.R. Brown, A. Campargue, J.-P. Champion, K. Chance, L.H. Coudert, V. Dana, V.M. Devi, S. Fally, J.-M. Flaud, R.R. Gamache, A. Goldman, etc, The HITRAN 2008 molecular spectroscopic database. // Journal of Quantitative Spectroscopy and Radiation Transfer, 2009, Volume 110, Issue 9, Pages 533-572, DOI: 10.1016/j.jqsrt.2009.02.013.

Рис. 1. Интерфейс для выбора источника данных с целью его декомпозиции по первичным источникам данных



2009_RoGoBaBe_HDO	L.S. Rothman, I.E. Gordon, A. Barbe, D.Chris Benner, P.F. Bernath, M. Birk, V. Boudon, L.R. Brown, A. Campargue, J.-P. Champion, K. Chance, L.H. Coudert, V. Dana, V.M. Devi, S. Fally, J.-M. Flaud, R.R. Gamache, A. Goldman, etc, The HITRAN 2008 molecular spectroscopic database. // Journal of Quantitative Spectroscopy and Radiation Transfer, 2009, Volume 110, Issue 9, Pages 533-572, DOI: 10.1016/j.jqsrt.2009.02.013.

☑	**Ошибка определения частоты ($см^{-1}$)**	0.1
☑	**Ограничения на частоту ($см^{-1}$)**	0 - 20000
☑	**Декомпозиция по эксперименту**	
☐	**Декомпозиция по расчету**	

Провести декомпозицию

Показать 40 строк от 0 | Всего строк 72 | > >> | Настройки

Название Вычисления/Эксперимент	Число записей	Число совпадений [z-a]	Показать	Публикация
2007_VoNaCaCO_HDO	3481	3330	Показать	B.A. Voronin, O.V. Naumenko, M. Carleer, P.-F. Coheur, S. Fally, A. Jenouvrier, R.N. Tolchenov, A.C. Vandaele and J. Tennyson, HDO absorption spectrum above 11 500 cm^{-1}: Assignment and dynamics. // Journal of Molecular Spectroscopy, 2007, v. 244, no. 1, p. 87-101.
2003_JaTeBeZo_HDO	9896	1592	Показать	A. Janca, K. Tereszchuk, P. F. Bernath, N. F. Zobov, S. V. Shirin, O. L. Polynsky, J. Tennyson, Emission spectrum of hot HDO below 4000 cm^{-1}. // Journal of Molecular Spectroscopy, 2003, v. 219, p. 132.
2007_JeDaReTy_HDO	3088	1463	Показать	Alain Jenouvrier, Ludovic Daumont, Laurence Regalia-Jarlot, Vladimir G. Tyuterev, Michel Carleer, Ann Carine Vandaele, Semen Mikhailenko, Sophie Fally, Fourier transform measurements of water vapor line parameters in the 4200–6600 cm^{-1} region. // Journal of Quantitative Spectroscopy and Radiation Transfer, 2007, v. 105, p. 326-355.
2007_JeDaReTy_HDO	3086	1463	Показать	Alain Jenouvrier, Ludovic Daumont, Laurence Regalia-Jarlot, Vladimir G. Tyuterev, Michel Carleer, Ann Carine Vandaele, Semen Mikhailenko, Sophie Fally, Fourier transform measurements of water vapor line parameters in the 4200–6600 cm^{-1} region. // Journal of Quantitative Spectroscopy and Radiation Transfer, 2007, v. 105, p. 326-355.
1982_PaCaFlGu_HDO	1902	1308	Показать	N. Papineau, C. Camy-Peyret, J. -M. Flaud and G. Guelachvili, The $2\nu_2$ and ν_1 bands of $HD^{16}O$. // Journal of Molecular Spectroscopy, 1982, v. 92, no. 2, p. 451-468.

Рис. 2. Результат декомпозиции данных об изотопомере HDO из банка данных HITRAN с точностью 0.1 $см^{-1}$ при разложении по расчетным и экспериментальным данным в диапазоне частот переходов 0-20000 $см^{-1}$

Таблица 2. Декомпозиция данных HITRAN по изотопомеру воды (HDO), выполненная по полному набору источников данных, собранному группой данных IUPAC [17, 18]

Частотный интервал	$N_{Hi-tran}$	ε	Источники данных из полного набора [6]	Остаток
0 – 10 cm^{-1}	65	10^{-1} cm^{-1}	**1946**_ToMe**, 1948**_StWeHiWa**, 1949**_Strandbe**, 1949**_Jen**, 1953**_BuSt, **1953**_BeWe**, 1953**_PoSt**, 1953**_JeBiMa**, 1955**_WeBeHe**, 1956**_ErCo**, 1957**_Posener**, 1962**_TrBe**, 1964**_ThKrLo**, 1968**_VeBlDy**, 1967**_BlVeDy**, 1970**_BeSt**, 1970**_StBe**, 1971**_LuCoHeGo**, 1973**_ClBeKlRo**, 1985**_Johns**, 1993**_GoFeDeDu	28
		10^{-1} cm^{-1}	***2000***_*ScPa 2007_ScPaTa_b, 2007_ScPaTa_a*	8
10 – 30 cm^{-1}	76	10^{-3} cm^{-1}	**1971**_LuCoHeGo**, 1976**_FlGi**, 1984**_MeLuHe**, 1985**_Johns	39
		10^{-2} cm^{-1}	***2000***_*ScPa, 2007_ScPaTa_b, 2007_ScPaTa_a*	38
30 – 50 cm^{-1}	76	10^{-2} cm^{-1}	**1976**_FlGi**, 1984**_MeLuHe, **1985**_Johns	61
		10^{-2} cm^{-1}	***2000***_*ScPa, 2007_ScPaTa_b, 2007_ScPaTa_a*	31
50 – 200cm^{-1}	599	10^{-1} cm^{-1}	**1978**_KaKaKy**, 1985**_Johns**, 1995**_PaHo	370
		10^{-1} cm^{-1}	***2000***_*ScPa, 2007_ScPaTa_b, 2007_ScPaTa_a*	0
200 –10000 cm^{-1}	8825	10^{-1} cm^{-1}	**1956**_BeGaPl, **1973**_CaFlGuAm, **1978**_KaKaKy, **1982**_PaCaFlGu, **1982**_ToGuBr, **1983**_Guelashv, **1983**_ToBr, **1985**_Johns, **1986**_FlCaMaGu, **1989**_OhSa, **1991**_SaTaIrNa, **1991**_RiSmDeBe, **1991**_RiSmMeBe, **1992**_RiSmDeBe, **1993**_Toth, **1995**_PaHo, **1997**_Toth_a, **1997**_Toth_b, **1999**_Toth, **2000**_WaHeHuZh, **2000**_SiBeMaMa, **2001**_PaBeZoSh, **2003**_JaTeBeZo, **2003**_BeNaCa, **2004**_NaVoHu, **2005**_ToNaZoSh, **2005**_ToTe, **2007**_JeDaReTy, **2007**_JeDaReTy, **2007**_MiLeKaCa, **2008**_*Guelashv_calib*	1715
		10^{-1} cm^{-1}	***2000***_*ScPa, 2007_ScPaTa_b, 2007_ScPaTa_a*	11
10000 – 20000 cm^{-1}	3483	10^{-1} cm^{-1}	**1997**_VoFaPlRiNe, **1998**_LaPeSiZh, **1999**_NaBeCa, **2000**_NaCA, **2000**_CaBeNa, **2000**_NaBeCa, **2000**_NaBeCaSc, **2000**_BeNaCa, **2001**_JeMeCaCo, **2004**_NaHuHeCa, **2005**_CaVaNa, **2005**_ToNaZoSh, **2007**_VoNaCaCO, **2008**_NaVoMaTe	8
		10^{-0} cm^{-1}	***2000***_*ScPa, 2007_ScPaTa_b, 2007_ScPaTa_a*	287

Заметим, что общее число публикаций, собранных группой IUPAC по изотопомеру молекулы HDO, составляет около 90 статей. В первой колонке таблицы указан частотный интервал, для которого проводилось разложение. Сделано это по причине зависимости точности измерений от величины частоты. Наиболее точные измерения (до 9 знаков) проводятся в микроволновом диапазоне (0 –10 $см^{-1}$). Во второй колонке указано число переходов из экспертного массива данных HITRAN, попадающих в данный частотный диапазон. В третьей колонке указана величина ε, входящая в формулу (1), указывающая на точность, с которой проводилась декомпозиция. В четвертой колонке дан перечень статей, в которых найдены частоты с заданной точностью. Мы не приводим библиографию в виду ее обширности и отсылаем читателя к работе [23] в которой дано подробное описание коллекции данных. Отметим, что теоретические списки переходов взяты из работ 2000_ScPa, 2007_ScPaTa_b, 2007_ScPaTa_a. Наконец, в последней колонке находится число переходов, не содержащихся в полном наборе публикаций.

Отметим, что разложение части банка данных HITRAN, относящейся к основному изотопомеру сероводорода, дает результат еще более несуразный. Из более чем 35 опубликованных работ по переходам в банке данных HITRAN используются данные из 11 работ, причем 70% данных из банка данных HITRAN не опубликованы [24].

Включение неопубликованных данных в экспертные массивы данных не является необычным фактом. Решение об использовании такого механизма определяется соответствующим научным сообществом. Однако с точки зрения автоматической обработки данных агентами невозможность найти данные в информационном пространстве будет означать их «несуществование». Именно в таком аспекте и рассмотрен критерий существования данных в нашей работе.

5 Заключение

В работе дано краткое описание задачи вычисления достоверности информационных ресурсов, публикуемых центром данных. Решение этой задачи использовано для создания прототипа инструментария публикации для виртуального центра молекулярных данных, созданного в рамках информационной системы W@DIS (http://wadis.saga.iao.ru). Дальнейшее развитие созданного инструмента публика-

ций будет осуществляться в рамках европейского инфраструктурного проекта VAMDC (Virtual Atomic and Molecular Data Center) как для атомных, так и молекулярных данных для широкого круга задач спектроскопии. Оно потребует детализации количественных ограничений на точность измерений по спектральным интервалам.

Литература

[1] NASA Langley Research Center (Radiation Budget, Clouds, Aerosols, Tropospheric Chemistry). – http://eosweb.larc.nasa.gov/.

[2] Earth Resources Observation and Science (EROS) Center. – http://eros.usgs.gov/.

[3] Solar Influences Data Analysis Center (SIDC). – http://sidc.oma.be/.

[4] The British Atmospheric Data Center (BADC). – http://badc.nerc.ac.uk/home/index.html.

[5] Data Center for Astrophysics. – http://www.isdc.unige.ch/.

[6] Atomic Mass Data Center. – http://amdc.in2p3.fr/.

[7] Virtual Atomic and Molecular Data Center. – http://vamdc.eu/.

[8] Реестр проекта VAMDC. – http://registry.vamdc.eu/vamdc_registry/main.

[9] Информационная система W@DIS. – http://wadis. saga.iao.ru.

[10] XML Schema for Atoms, Molecules and Solids (XSAMS). – http://www-amdis.iaea.org/xsams.

[11] Voronin B.A., Tennyson J., Tolchenov R.N. et al. A high accuracy computed line list for the HDO molecule // Monthly Notices of the Royal Astronomical Society. – 2010. – V. 402. – P. 492-496.

[12] Public Astronomical Catalogues and Lists. – ftp://cdsarc.u-strasbg.fr/pub/cats/.

[13] De Roure D., Jennings N., Shadbolt N. A future e-science infrastructure // Report Commissioned for EPSRC/DTI Core e-Science Programme, 2001. – 78 p.

[14] WWW Corporation. – http://w3c.org/ standards/.

[15] Pietruszczak A. Pieces of mereology// Logic and Logical Philosophy. – 2005. – V. 14. – P. 211-234.

[16] Зиновьев А.А. Основы логической теории знаний. – М.: Наука, 1967. – 260 с.

[17] RFC 2396, Uniform Resource Identifiers. – http://www.ietf.org/rfc/rfc2396.txt.

[18] Быков А.Д., Науменко О.В., Родимова О.Б. и др., Информационные аспекты молекулярной спектроскопии. – Томск, Изд-во ИОА СО РАН, 2008. – 256 с.

[19] Привезенцев А.И., Организация онтологических баз знаний и программное обеспечение для описания информационных ресурсов в молекулярной спектроскопии. – Дисс. … канд. техн. наук. – Томск, 2009. – 238 с.

[20] Privezentsev A., Fazliev A., Tsarkov D., Tennyson, J. Computed kknowledge base for description of information resources of water spectroscopy. – http://www.webont.org/owled/2010/.

[21] Rothman L.S., Gordon I.E., Barbe A. et al. The HITRAN 2008 molecular spectroscopic database// J. Quant. Spectr. Rad. Transfer. – 2009. – V. 110. – P. 533-535.

[22] Tennyson J., Bernath P.F., Brown L.R. et al. 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. – V. 110. – P. 573-596.

[23] Tennyson J., Bernath P.F., Brown L.R. et al. IUPAC critical evaluation of the rotational-vibrational spectra of water vapor. Part II. Energy levels and transition wavenumbers for $HD^{16}O$, $HD^{17}O$ and $HD^{18}O$// J. Quant. Spectr. Rad. Transfer. – 2010 (to appear).

[24] Naumenko O.V., Brown L.R., Campargue A. et al. Critical evaluation of the vibrational-rotational transitions of hydrogen sulphide and its isotopologues from 0 to 16500 cm^{-1}// Proc. of 11-th HITRAN Database Conference, 2010. – P. 76.

Data and metadata publishing tools for a distributed information system on quantitative spectroscopy

A.Yu. Akhlyostin, N.A. Lavrentiev, M.M. Makogon, A.I. Privezentsev, A.Z. Fazliev

In our report some features of publication tools, partially developed by our group in the framework of VAMDC project, are discussed. Two of them, namely, data validity and data manipulation, are the main topic of the report.

The following two types of constraints are discussed. The first one is the restrictions on the values of physical quantities derived from the mathematical model of processes, in particular, selection rules. The results of the verification of more than 700 primary data sources related to water spectroscopy are discussed.

Another type of restrictions relate to the fact of data publication (existence constraint). The results of the checkup applied to Hitran data for a series of water isotopomers are discussed. These checkups are based on the data collected by IUPAC group. The problem of creation of an information system containing a complete set of published data for a series of atmospheric molecules is discussed. The state of the art of the problem developed in IAO SB RAS is described.

♣ Работа выполнена при финансовой поддержке РФФИ (проект 08-07-00318) и 7-й рамочной программы ЕС (грант 239108)