ИНСТРУМЕНТ ПУБЛИКАЦИИ В ПРОЕКТЕ «ВИРТУАЛЬНЫЙ ЦЕНТР АТОМАРНЫХ И МОЛЕКУЛЯРНЫХ ДАННЫХ»

Фазлиев А.З., Привезенцев А.И., Ахлёстин А.Ю., Лаврентьев Н.А., Козодоев А.В.
Институт оптики атмосферы имени В.Е. Зуева СО РАН (Томск), Россия
faz@iao.ru

Аннотация

В докладе дано краткое описание проекта седьмой рамочной программы «Виртуальный центр атомных и молекулярных данных». Особое внимание уделено задаче создания инструмента проверки качества данных. Решение этой задачи использовано для создания инструментария публикации в информационной системы W@DIS. Дальнейшее развитие созданного инструмента публикаций 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.

1. ВВЕДЕНИЕ

Научное исследование требует усвоения найденного и полученного. Цель работы большинства ученых состоит в том, чтобы другие люди (ученые, инженеры, чиновники или общественность) могли понять и использовать результаты их работы. Наука является наиболее интернациональной человеческой деятельностью, расцветающей при международном обмене на любых уровнях. Такие обмены традиционны, важны и продуктивны, в силу чего национальные границы становятся прозрачными во время контактов ученых. В настоящее время значительная часть обмена имеет цифровую форму, в которой огромные массивы данных сформированы во всех формах современных информационных технологий. Они заставляют ученых приспосабливаться в процессе исследований к данным, объемы которых быстро увеличиваются и процедурам их усвоения, часто с целями междисциплинарного взаимодействия в формах неизвестных ранее. Стиль изменения данных порождает изменения в мировых "центрах данных", организациях производителей данных и пользователей, являющихся участниками процессов таких революционных преобразований. Предпочтительной формой для общения в таких условиях является World Wide Web, благодаря его повсеместности и демократичности.

Текущее состояние дел с хранением и оперированием данными включает традиционные проблемы согласованности наборов данных: обеспечение исследователей способами нахождения доступных компиляций данных, обеспечения достаточного количества описательных материалов (метаданных), дающих исследователям возможность использовать базы данных сформированных с ограничениями целостности данных. Недавнее быстрое увеличение количества данных в науках, использующих наблюдения (астрономия), благодаря автоматическим методам сбора данных и их компиляции, высветило проблемы передачи, хранения и анализа данных.

Поскольку сказанное относится ко всем наукам, оно в частности, проявилось в области атомной и молекулярной науки. Эти предметные области порождают данные, использующиеся в широком спектре научных и технических приложений. Прогресс в научных и технических областях возможен лишь при точной количественной информации по свойствам столкновений и спектроскопическим характеристикам взаимодействующих веществ. Атомные и молекулярные данные необходимы в таких несхожих приложениях как астрофизика, атмосферные науки, дисциплинах, изучающих энергию расщепления, в производстве полупроводников и других, использующих плазму, технологиях, the lighting industry, при обнаружении и удалении загрязнений (прежде всего, для обнаружения взрывчатых и биологических агентов, применяемых террористами), и являются существенными для понимания многих биологических процессов, включая моделирование радиационного поражения в клеточных системах для последующего терапевтического вмешательства. Ученые, работающие с атомными и молекулярными данными, обеспечивают основу для новой эры исследований - e-Sciences.

Тем не менее признается, что остается несколько проблем развития надежной и интегрированной инфраструктуры, пригодной для использования пользователями. Существующие проблемы можно разделить на две категории: (1) полнота и качество данных и (2) интерфейсы для интеллектуального анализа данных. Сегодня такие задачи решаются во многих центрах данных, и решения сфокусированы на специальных приложениях и не являются гибкими. Следовательно, насущно необходимо:

- *развивать тесные связи между пользователями, производителями данных и центрами данных и внедрять современные технологии.*
- *устанавливать лучшую международную кооперацию для продвижения компиляций атомных и молекулярных данных и работ по базам данных, избегать дублирования усилий и гарантировать использование наиболее приемлемых данных.*

Многие европейские группы признаны за их превосходство в области атомных и молекулярных наук и вклад в измерения и вычисления атомных и молекулярных данных, положенных в основу нескольких баз данных и сервисов. В некоторых случаях, производители данных уже сформировали сети европейского масштаба, например, European Theoretical Spectroscopy Facility [1], проект поддержанный Nanoquanta Network of Excellence [2].

Некоторые базовые инициативы ERA планируют использование таких сервисов для собственных целей и нужд (например, планетарные науки с помощью Europlanet I3-FP7, астрономы с помощью Euro-VO I3-FP7, сообщество, исследующее расщепление (поддержано ITER и EURATOM) и сообщество радиационных наук для моделирования радиотерапии и влияния малых доз на здоровье человека (поддержано EURATOM и Framework VII (Health) Programme). Однако в рамках этих инициатив слишком мала финансовая поддержка на работу, связанную с интероперабельностью атомных и молекулярных данных. Каждый раз одни и те же базы данных используются для нового приложения, а выходные массивы из баз данных адаптируются к решаемым задачам. Например, планируемый автоматический инструмент для визуализации планетарной, звездной или межзвездной среды. Эти инструменты требуют автоматического доступа к разным атомно-молекулярным базам данных, сверке извлекаемых данных, проверке качества данных. На сегодняшний день нет общей инфраструктуры для решения таких задач.

Целью Virtual Atomic and Molecular Data Centre (VAMDC) является построение такой безопасной, документированной, гибкой, легко доступной и интероперабельной цифровой инфраструктуры для атомарных и молекулярных данных. VAMDC будет построен исходя из экспертизы существующих баз данных, производителей данных и производителей сервисов с целью создания инфраструктуры которая, с одной стороны, позволит извлекать данные из существующих хранилищ и, с другой стороны, достаточно гибко приспособлять их к нуждам широкого круга пользователей из академического и промышленного сообщества или из общественности, как в рамках ERA, так и вне их. Проект нацелен на построение ядра сообщества, развертывание инфраструктуры и разработку программного обеспечения, а также на обеспечение обучения потенциальных пользователей и усвоение посредством ERA. Предполагается, что VAMDC станет легальным европейским объектом в результате выполнения проекта.

Ключевым моментом для достижения такой цели является преодоление фрагментации в сообществе исследователей, занимающихся атомарными и молекулярными базами данных. VAMDC достигнет этого посредством:

- развития самой большой и наиболее представительной разделяемой атомной и молекулярной инфраструктуры, поддерживаемой и расширяемой всеми учеными европейского союза
- обеспечением распределенной европейской инфраструктуры доступной, используемой и упоминаемой широким европейским исследовательским сообществом.

Доклад основан на материалах гранта FP7 VAMDC [3] и статьи [4]. Оригинальной частью доклада являются разделы 5 и 6.

2. ИНФОРМАЦИОННЫЕ РЕСУРСЫ УЧАСТНИКОВ ПРОЕКТА VAMDC

Проект VAMDC объединил производителей спектральных данных и разработчиков информационных систем из более чем 20 организаций трех континентов. Участники проекта содержат ресурсы, относящиеся к разным областям спектроскопии и, частично, химии. Созданные ими ресурсы отличаются по структуре, наполнению данными и фактами, качеству данных и ряду других характеристик. Они также сильно различаются по типам доступа, способам представления данных, наличию метаданных, и других свойствам информационных ресурсов. Некоторые детали этих ресурсов описаны в [4], содержащей также ссылки на авторское описание ресурсов. Здесь мы ограничимся их кратким перечнем.

Ресурсы, связанные с атомной спектроскопии, включают в себя

1. базу данных VALD спектральных параметров атомов [5],
2. базу данных CHIANTI, содержащую информацию о спектральных свойствах ионов [6],
3. базу данных Stark-B, содержащую вычисленные полуширины и сдвиги изолированных спектральных линий атомов и ионов обусловленные столкновениями с электронами и ионами [7],
4. базу данных Spectr-W3, содержащую данные о спектральных характеристиках атомов и ионов [8],
5. базы данных IVIC/CeCalCULA (TIPTOPbase, OPserver, XSTAR), содержащие атомные данные [9-11],

6. базу данных о спектральных свойствах атомов, созданную в NIST и содержащую около 77,000 уровней энергии, 144000 спектральных линий 99 элементов таблицы Менделеева и ряда изотопов [12].

Ресурсы, связанные с молекулярной спектроскопией газов, включают в себя

1. базу данных EMol, содержащую списки измеренных и вычисленных сечений, характеризующих взаимодействие электронов с молекулярными системами [13],
2. базы данных ИОА CO РАН, содержащие спектральные характеристики диоксида углерода и озона, информационные системы [14, 15], включающие в себя экспертные массивы данных HITRAN [16], GEISA [17] и HITEMP [18], первичные источники данных о спектральных характеристиках ряде атмосферных молекул (W@DIS, [19]),
3. базу данных Cagliari, содержащую спектральные характеристики ароматических полициклических углеводородов и углеродных кластеров в четырех состояниях, описываемых зарядом: анион, катион, нейтральный и дикатион [20],
4. базу данных BASECOL, содержащую коэффициенты для вычисления скорости возбуждения необходимые при описании процессов колебательно-вращательного возбуждения молекул электронами [21].

Ресурсы, связанные с молекулярной спектроскопией вещества в твердой или жидкой фазе, включают

1. базу данных GhoSST [22], содержащую спектральные данные о веществе молекулярного или атомного состава, находящемся в твердом или жидком состоянии, относящиеся к инфракрасному, видимому и ультрафиолетовому диапазону,
2. базу данных LASP [23], содержащую ИК спектры молекул в твердом состоянии (T=10–100K) как для отдельных молекул, так и для их смесей до и после их обработки ионами с энергиями (30–200keV) и фотонами (10.2eV), оптические константы молекул в твердой фазе в ИК диапазоне, оптические константы замороженных молекул после обработки ионами энергий (30–200 keV), интенсивности полос наиболее представительных полос ИК диапазона и значения плотностей замороженных образцов.

Ресурсы, связанные с химическими процессами включают в себя

1. базу данных UMIST [24], содержащую данные о скоростях химических реакций актуальных для астрохимии,
2. базу данных KIDA [25], содержащую химические реакции, используемые при моделировании химических процессов в межзвездной среде и планетарных атмосферах.

Специализированная база данных CDMS [26] содержит рекомендуемые значения частот переходов и интенсивностей для атомов и молекул в диапазоне 0–340 cm^{-1}.

Из перечня ресурсов можно сделать вывод о том, что эти данные ориентированы на разные группы прикладных задач в таких предметных областях как астрономия (включая химические процессы в межзвездном пространстве и атмосферах экзопланет), оптика атмосферы, диагностике плазмы, прикладных задачах радиационных науках (от использования радиации в медицине, до исследования процессов переноса радиации в средах). Большая часть ресурсов содержится в базах данных, но небольшая часть ресурсов хранится в файловых системах. В настоящее время научные информационные системы в

спектроскопии на практике являются мало используемыми в виду трудоемкости их создания. Заметим, что мы не относим к числу научных информационных систем системы, содержащие базы данных или файловые системы и связанные с ними интерфейсы.

Выделим ключевые особенности ресурсов, попавших в проект VAMDC. Во-первых, число сущностей, характеризующих интенсионал данных существенно меньше числа фактов, входящих в экстенсионал данных и большая часть фактов содержит числовые значения. Такое отношение характерно для точных наук, в которых количественный аспект доминирует. Во-вторых, в перечисленных выше предметных областях спектральные данные используются в качестве входных данных в задачах этих предметных областей. Проблема идентичности интенсионалов данных не является определяющей при моделировании, т.к. взаимодействующие между собой исследователи в спектроскопии и прикладных науках в той или иной мере решили проблему качества понятий близких по семантике. Это обстоятельство позволяет построить согласованную модель данных для этих предметных областей. В-третьих, для количественных областей разработаны изящные, но иногда трудоемкие, методы проверки качества данных, позволяющие автоматизировать процесс контроля достоверности данных.

3. XSAMS – XML-СХЕМА ДЛЯ ОПИСАНИЯ СТРУКТУРЫ ДАННЫХ, ОТНОСЯЩИХСЯ К ПРОЦЕССАМ ВЗАИМОДЕЙСТВИЯ ИЗЛУЧЕНИЯ С АТОМАМИ, МОЛЕКУЛАМИ И ПОВЕРХНОСТЬЮ ТВЕРДЫХ ТЕЛ

Основное назначение схемы данных XSAMS [27] состоит в обеспечении корректного обмена данными между организациями. Стоит отметить, что даже простые проверки данных, допустимые стандартом XML, а тем более проверка качества данных, не являются целевой функцией этой схемы. Хотя XSAMS требует включения библиографических ссылок или идентификаторов ресурсов вопросы проверки корректности данных остаются за их производителем или хранителем. Другими словами, схема дает форму, в которой осуществляется обмен данными, но оставляет за поставщиком данных вопросы соответствия значений данных интенсионалу данных и проверки ограничений на значения данных. Это означает, что в результате обмена пользователь может получить такие

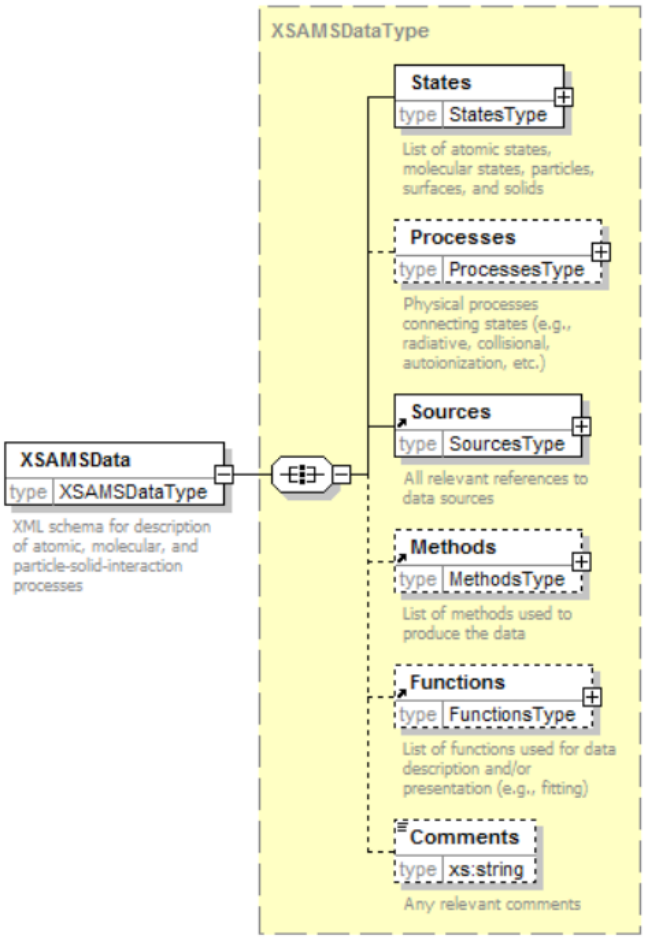

Рис.1 Основной уровень структуры данных в XSAMS [http://www-amdis.iaea.org/xsams/docu/v0.1.pdf].

экзотические наборы данные, которые могут содержать отрицательные интенсивности, частоты и значения квантовых чисел.

Структура XSAMS состоит из описания ряда физических процессов, описанных в терминах состояний. Состояния атомов, молекул, ионов и твердых тел описывается как можно детальнее. Процессы между состояниями описываются путем ссылки на начальное и конечное состояние. Состояние специфицируется с любым количеством деталей и в разных схемах связи.

4. РАБОЧИЕ ПРОГРАММЫ VAMDC

Проект содержит 8 рабочих программ, отнесенных к трем группам деятельности. Первая группа «Сетевая деятельность» включает в себя три программы, обеспечивающие управление проектом, научно-техническую координацию, внедрение результатов проета и обучение пользователей работе с созданном в проекте инструментарием.

Вторая группа «Сервисная деятельность» содержит две программы, обеспечивающие развитие и поддержку инфраструктуры.

Третья группа «Совместная исследовательская деятельность» включает в себя три рабочие программы: «Интероперабельность», «Инструмент публикации» и «Инструменты анализа данных и интеграции».

Рассмотрим некоторые детали последних трех рабочих программ.
Программа «Интероперабельность» содержит четыре задачи.

1. *Модели данных и XML схема*. Решение этой задачи приведет к построению моделей данных и XML-схем для данных, относящихся к описанию физических процессов (ассоциации, диссоциации, высвечивания, уширения и сдвигов спектральных линий, упругого рассеяния, переноса заряда, рекомбинации, флуоресценции, ионизации и т.д.), в которых участвуют атомы, молекулы ионы, изотопы, ароматические полициклические углеводороды, а также процессы на поверхности твердых тел.
2. *Словари*. Решение этой задачи должно обеспечить единые обозначения для веществ, кодировку физических и химических процессов, систематизацию квантовых чисел характеризующих состояния атомов и молекул.
3. *Протоколы доступа и языки запросов и извлечения*. Должны быть определены протоколы запросов для извлечения разных типов ресурсов: численных данных, ссылок и документов.
4. *Реестры* [28]. Должны быть составлены реестры, с помощью которых приложения смогут находить и извлекать ресурсы, т.е. данные и сервисы, необходимые для решения научных проблем.

Программа «Инструмент публикации» содержит пять задач. Эта программа ориентирована на создание инструмента являющегося часть инфраструктуры проекта VAMDC. Она должна следовать стандартам, разработанным в предыдущей программе. С помощью этого инструмента пользователи будут обращаться через технический сайт проекта к ресурсам участников проекта.

1. *Развитие интерфейса для работы с базами данных участников проекта*. Основное требование к создаваемому интерфейсу – это обеспечение доставки данных

в формате XSAMS в базы данных и экспорт из баз данных запрашиваемых данных в формате XSAMS.

2. *Построение и модернизация реестра в части содержания баз данных*. Создание сервисов обеспечивающих актуализацию реестра информационных ресурсов Виртуального центра атомных и молекулярных данных.

3. *Инструмент импорта данных*. Развитие, документирование и тестирование инструмента импорта новых данных.

4. *Графический интерфейс пользователя для работы с публикуемыми данными*. Создание интерфейса, обеспечивающего формирование запросов пользователем.

5. *Инструмент автоматического контроля качества данных.* Создание и тестирование программного обеспечения, обеспечивающего проверку ограничений на значения данных.

Программа «Инструменты анализа данных и интеграции» состоит из трех задач.

1. *Запросы к реестру ресурсов*. Задача состоит в нахождении ресурсов по некоторым ключевым словам в их описании. Необходимо определить и реализовать протокол запросов к реестрам.

2. *Инструмент манипуляции данными*. Инструмент для объединения или пересечения приходящих по запросу данных.

3. *Сервисы извлечения шаблонов структур из данных (data mining) с целью создания новых данных*. Задача состоит в выработке стратегии формирования атомно-молекулярных ресурсов из имеющихся в наличии с целью обеспечения данных некоторых выделенных групп пользователей (например, астрономов, получающих спутниковые данные). Необходимо создать сервисы создающие сложные комбинации таких работ как доступ к данным, манипуляция, интеграция данных в задачи пользователя.

5. ПРОЦЕДУРА ПУБЛИКАЦИИ ДАННЫХ

В настоящее время не решен окончательно вопрос о процедуре публикации в части редактирования и рецензирования данных в Виртуальном центре. Обсуждаются несколько вариантов. Наиболее простой из них сформулирован в предложении отказаться от рецензирования и размещать в базах данных участников данные, которые они считают нужными. Это предложение облегчает работу создателей инструмента публикации. Вся экспертная работа в этом случае выполняется исследователями, поддерживающими базы данных. Предполагается, что они более детально знают задачи, для которых пользователям необходимы данные. Другое предложение состоит в использовании процедуры рецензирования на начальном уровне представления производителем данных. В этом случае технической сложностью является организационная процедура выбора рецензентов. Наконец, третье предложение состоит в организации автоматической проверки формальных ограничений следующих из математических моделей атомов и молекул и процессов, в которых они принимают участие и ряда общих ограничений связанных с отношениями рефлексивности и транзитивности. Эта проверка должна осуществляться на техническом сайте проекта, через который идут потоки запросов и импорт данных. Каждое из этих трех

предложений. Преимущество последнего предложения состоит в том, что значительный объем проверок производитель может выполнить самостоятельно, используя инструментарий для проверки формальных ограничений на техническом сайте проекта. При такой процедуре на долю рецензента выпадает проверка не формальных ограничений, например, связанных с условиями экспериментов или корректностью математических моделей в расчетах.

Необходимость создания инструментария для проверки формальных ограничений и обеспечения работы экспертов с данными производителей не зависит от того, какое из предложений будет принято. Однако формальные критерии для проверки ограничений в значительной степени зависят от структур данных имеющихся у участников проекта. В первую очередь это относится к проверке на существование информационного ресурса. В чем состоит необходимость такой проверки? Подавляющее большинство спектроскопических ресурсов представляет собой составные массивы данных, содержащие части, которые могут быть не опубликованными.

Построение экспертного массива спектральных данных состоит в составлении канторова множества (множества в котором все элементы уникальны). В силу того, что измерения проводятся с некоторой точностью, то фактический интерес представляет мультимножество [29] переходов, вычисленных при обработке измерений. Исследователи, к сожалению, в базах данных размещают только экспертные массивы, а не данные из публикаций (мультимножества). Среди участников проекта VAMDC только в ИОА СО РАН в информационной системе W@DIS [30] можно генерировать мультимножества переходов для ряда молекул и их изотопомеров. По этой причине в настоящее время невозможно организовать проверку ограничения существования в рамках проекта VAMDC.

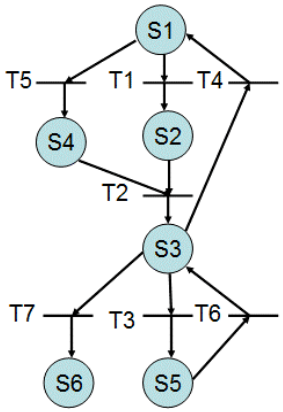

Fig. 2. Сеть Петри для описания процедуры публикации

Состояния
S1 – «Пользователь», S2 - «Пользователь обладает собственными данными в РИС», S3 – «Редактор»
S4 - «Пользователь модифицировал собственные данные», S5 – «Рецензент», S6 – «Опубликованный источник данных»
Переходы
T1 – создание пользователем нового источника данных в ИС
T2 – пересылка адреса нового источника редактору
T3 – пересылка данных рецензентам
T4 – посылка письма пользователю (данные отклонены или необходимо модифицировать данные
T5 – загрузка модифицированных данных
T6 – посылка рецензии редактору
T7 – публикация источника данных
Ограничения
Переход T1 – данные не были загружены в ИС
Переход T5 – данные загружены в ИС
Переход T4 – редактор передал данные рецензентам
Переход T7 – принято решение о публикации данных.

Рассмотрим каким образом публикация данных организована в распределенной информационной системе по спектроскопии, созданной в России авторами доклада. На рис.2 приведена сеть Петри, указывающая порядок работы с массивом данных при публикации. В

этой упрощенной сети есть 6 состояний, в которых может находиться массив данных. Состояние S1 соответствует всем ситуациям в которых у пользователя есть массив данных, но он не загружен в ИС. Состояние S2 соответствует ситуациям, когда пользователь загрузил данные в ИС, но еще не отправил их Редактору. Состояние S3 соответствует ситуациям, когда редактор получил данные от пользователя. Состояние S4 соответствует ситуации, когда пользователь повторно посылает данные на редактирование. Состояние S5 соответствует ситуациям, в которых массив данных рассматривается рецензентом. Состояние S6 характеризует данные принятые для публикации. Ввиду очевидной интерпретации сети Петри описание переходов и ограничений здесь опущено.

Загрузка данных в системе W@DIS осуществляется в несколько шагов. Для осуществления процедуры загрузки пользователь должен создать новый источник информации и присвоить ему название. Затем выбрать молекулу для которой он загружает данные, указать число сущностей которые используются в файле загрузке, дать описание интенсионала массива и указать место в котором находится загружаемый файл. На рис. 3 и 4

Показаны интерфейсы, используемые при загрузке.

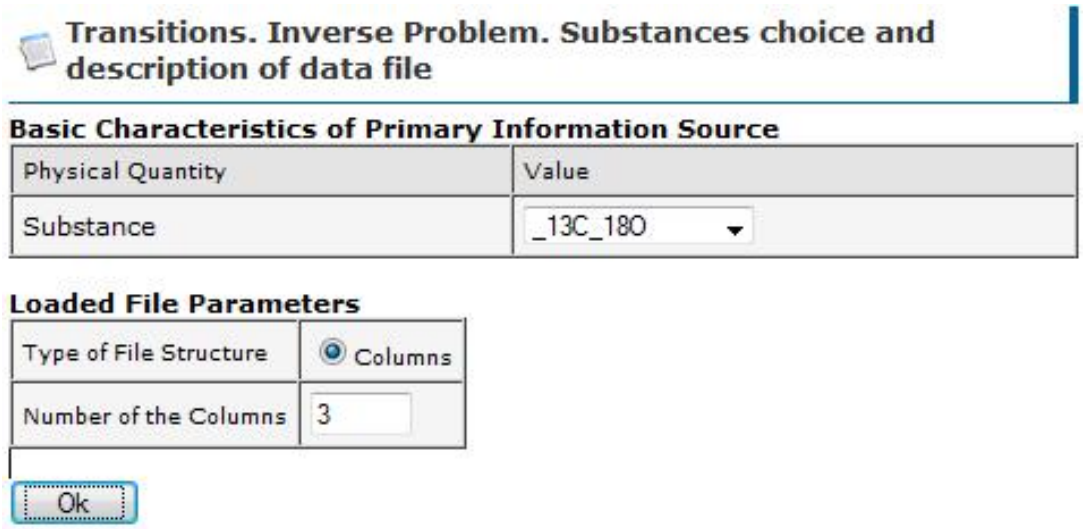

Рис.3. Первый шаг: Выбор молекулы и указание числа физических сущностей в загружаемом файле.

Для каждого загруженного в ИС массива данных система автоматически вычисляет значения метаданных и формирует индивид онтологии информационных ресурсов молекулярной спектроскопии.

Пользователь может просмотреть метаданные любого публичного массива данных и собственных данных. Пример представления метаданных показан на рис.5.

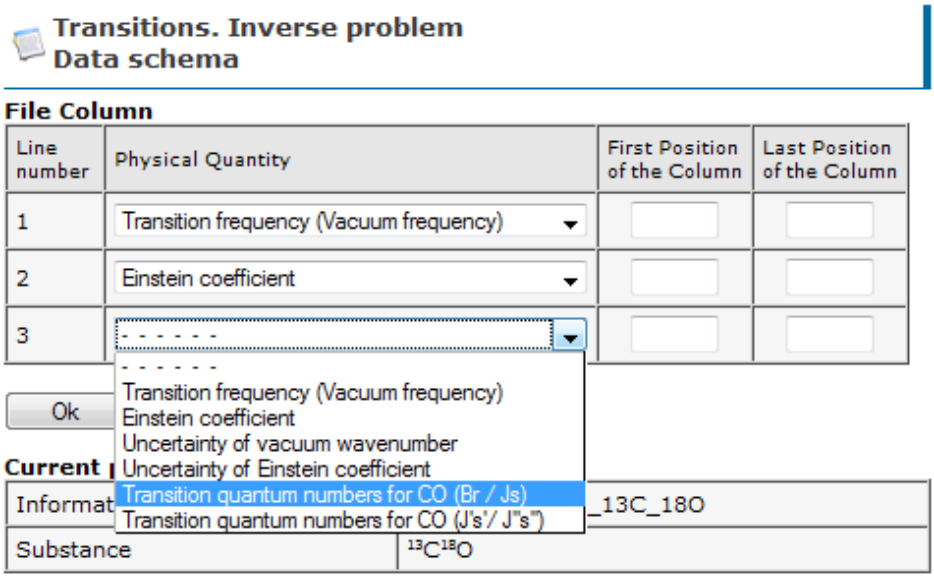

Рис.4. Второй и третий шаги: Построение схемы данных соответствующей данным размещенным в загружаемом файле и загрузка файла в ИС.

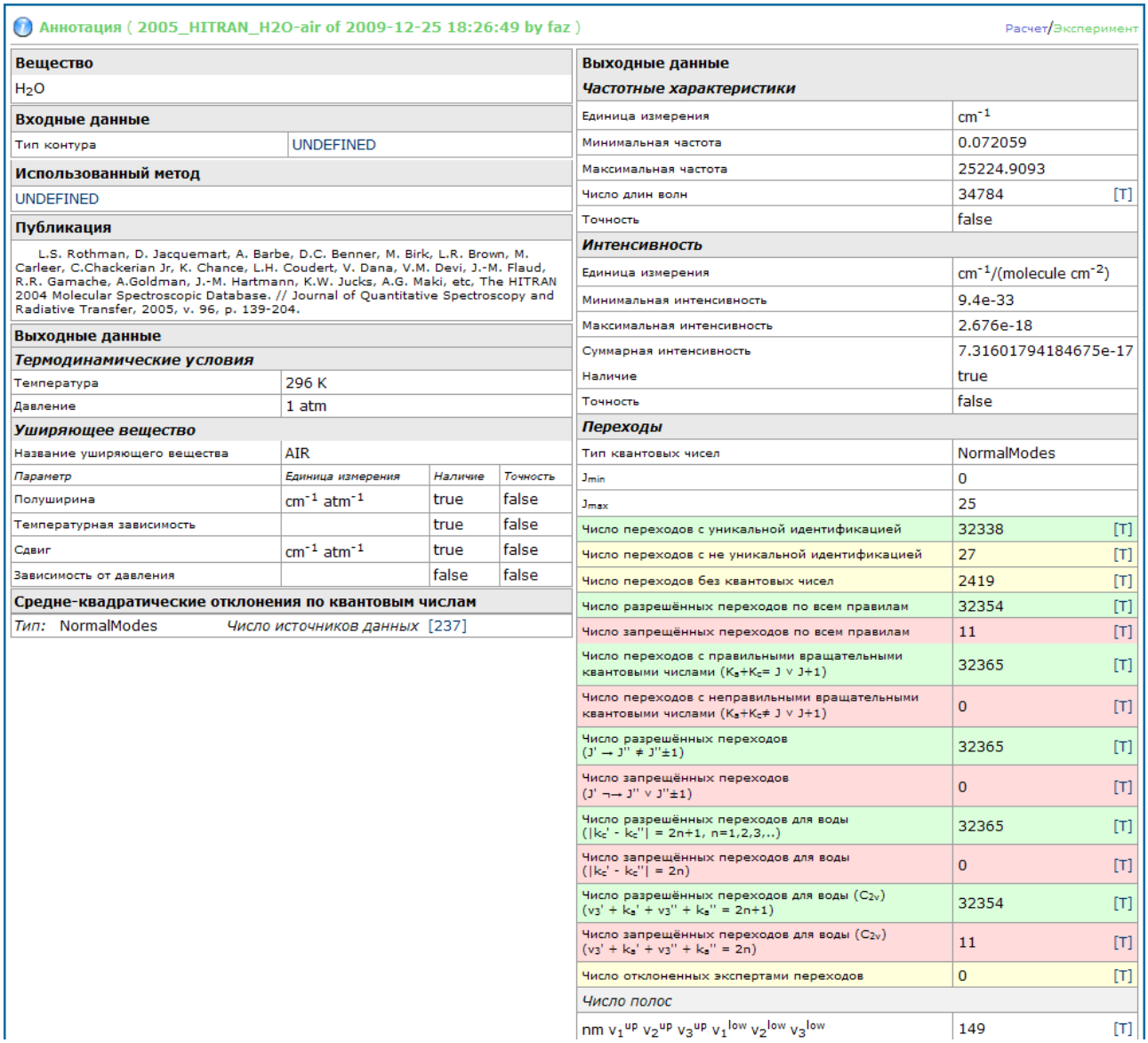

Рис.5. Метаданные для экспертного массива данных (HITRAN 2004).

Детали описания метаданных характерных для молекулярной спектроскопии можно найти в монографии [31] и диссертации [32].

Аннотации информационных ресурсов систематизированы в каталожной системе, позволяющей автоматически распределять ресурсы по каталогам. Основное внимание уделено образованию той части каталогов которая описывает достоверность первичных информационных ресурсов в молекулярной спектроскопии.

6. ДОСТОВЕРНОСТЬ ИНФОРМАЦИОННЫХ РЕСУРСОВ

6.1 Интерпретация достоверности

Определение достоверности данных в разных предметных областях разное [33]. Требования к данным в молекулярной спектроскопии намного жестче по сравнению с требованиями прикладных областей, в которых эти данные используются. Например, для вычисления поглощения потоков солнечного излучения квантовые числа, характеризующие переходы не обязательны, тогда как для вычисления уровней они необходимы. По этой причине программное обеспечение, реализующее проверку достоверности, будет разным для разных предметных областей.

Ниже проверка достоверности связана с проверкой ограничений на информационные ресурсы предметной области. Подробно рассмотрена группа ограничений на значения физических величин. Ограничения существования и их применения подробно разбираются в докладе Лаврентьева Н.А. и др. «Система проверки ограничений существования для молекулярной спектроскопии» в рамках данной конференции.

Ограничения на значения связаны 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 о том, что тот или иной переход относится к переходам рассматриваемой молекулы. При рассмотрении физических процессов к правилам отбора могут быть добавлены дополнительные ограничения или изменены правила отбора.

Вторая группа ограничений связана с интерпретацией существования информационных ресурсов [34], 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 о том что все используемые в экспертных массивах значений опубликованы по процедуре рекомендованной сообществом исследователей [35].

6.2 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. Она не связана с квантовыми числами.

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.

В табл. 1 приведены результаты классификации источников данных по спектроскопии сероводорода и углекислого газа 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. В Таб. 1 показаны результаты проверки все опубликованных данных в профильных журналах за период более чем 70 лет. В Таб.2 представлен анализ около 80% опубликованных работ. В этих таблицах не учтены результаты экспертных оценок, т.к. они в настоящее время не сделаны.

Таблица 1. Результаты проверки достоверности первичных источников информации о решениях задач спектроскопии сероводорода [36].

Вещество	Т1- Прямая задача определения уровней энергии	Т6 – Обратная задача определения вакуумных волновых чисел	Т7- Обратная задача определения уровней энергии
	NM (нормальные моды)	**NM**	**NM**
H_2S	9 (9)	22 (22)	19 (17)
HDS	5 (5)	5 (3)	6 (5)
D_2S	5 (5)	6 (6)	3 (3)
$H^{34}SH$	-	9 (7)	8 (8)
$H^{33}SH$	-	8 (7)	4 (4)
$HD^{34}S$	-	2 (2)	1 (1)
$D_2{}^{34}S$	-	1 (1)	-
Итого	19 (19)	53 (48)	41 (38)

Таблица 2. Результаты проверки достоверности источников информации о углекислом газе.

Вещество	**Т2** Прямая задача определения частот поглощения		**Т3** Прямая задача определения параметров контура линии		**Т5** Обратная задача определения параметров контура линии		**Т6** Обратная задача определения вакуумных волновых чисел	
	HN	**CN**	**HN**	**CN**	**HN**	**CN**	**HN**	**CN**
CO_2	23 (16)	4 (3)	16 (14)	1 (0)	56 (52)	-	54 (48)	-
$O^{13}CO$	12 (6)	4 (3)	7 (7)	1 (1)	22 (21)	-	27 (24)	-
$O^{13}C^{18}O$	4 (4)	2 (2)	2 (2)	-	11 (11)	-	14 (14)	-
$OC^{18}O$	12 (11)	4 (4)	5 (5)	1 (1)	14 (13)	-	22 (22)	-
$OC^{17}O$	10 (9)	4 (4)	5 (5)	1 (1)	11 (10)	-	17 (16)	-
$O^{13}C^{17}O$	3 (3)	2 (2)	-	-	6 (6)	-	8 (8)	-
$^{17}OC^{18}O$	1 (1)	-	1 (1)	-	5 (5)	-	5 (5)	-
$C^{18}O_2$	6 (4)	2 (2)	1 (1)	-	8 (7)	-	8 (7)	-
$^{13}C^{17}O_2$	1 (0)	-	-	-	-	-	1 (0)	-
$^{13}C^{17}O^{18}O$	1 (1)	-	-	-	-	-	3 (3)	-
$^{18}O^{13}C^{18}O$	4 (3)	-	-	-	3 (3)	-	7 (6)	-
$^{14}CO_2$	2 (1)	-	-	-	-	-	3 (3)	-
$C^{17}O_2$	-	-	1 (1)	-	1 (1)	-	1 (1)	-
$^{14}C^{18}O_2$	-	-	-	-	-	-	1 (1)	-
Всего	79 (59)	22 (20)	38 (36)	4 (3)	137 (129)	-	171 (158)	-

7. ЗАКЛЮЧЕНИЕ

В докладе дано краткое описание проекта седьмой рамочной программы «Виртуальный центр атомных и молекулярных данных». Особое внимание уделено задаче создания инструмента проверки качества данных. Решение этой задачи использовано для создания инструментария публикации в информационной системы W@DIS. Дальнейшее развитие созданного инструмента публикаций 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.

Отметим, что подход группы данных IUPAC [19, 36, 37] к систематизации всех опубликованных данных, с указанием некорректностей разного типа, встречающихся в опубликованных данных, а не с публикацией избранных экспертами данных, создал в спектроскопии качественной иной метод работы с информацией (данными, метаданными и онтологиями). Он существенно отличается от методологии работы со спектральными данными участников проекта VAMDC. Это означает, что по окончании проекта VAMDC придется заново переделывать созданное программное обеспечение, ориентировав его не только на работу с пользователем-человеком, но и пользователем-программным агентом или web-сервисом. Для решения этих задач потребуется создание онтологий информационных ресурсов по спектроскопии. Первый шаг в этом направлении уже сделан в работах [32, 38].

Литература

[1] European Theoretical Spectroscopy Facility, http://www.etsf.eu.

[2] Nanoquanta Network of Excellence , http://www.nanoquanta.eu.

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

[4] M.L. Dubernet, V. Boudon, J.L. Culhane, M.S. Dimitrijevic, A.Z. Fazliev, C. Joblin, F. Kupka, G. Leto, P. Le Sidaner, P.A. Loboda, H.E. Mason, N.J. Mason, C. Mendoza, G. Mulas, T.J. Millar, L.A. Nuñez, V.I. Perevalov, et al., Virtual atomic and molecular data centre, J. Quant. Spectr. Rad. Transfer, 2010, V. 111, Issue 15, P. 2151-2159.

[5] U. Heiter, P. Barklem, L. Fossati, R. Kildiyarova, O. Kochukhov, F. Kupka, M. Obbrugger, N. Piskunov, B. Plez, T. Ryabchikova, H. C. Stempels, Ch. Stutz and W. W. Weiss, VALD – an atomic and molecular database for Astrophysics, J. Phys.: Conference Series **130** (2008) 012011.

[6] Dere K.P., Landi E., Young P.R., Del Zanna G., Landini M., Mason H.E., CHIANTI—an atomic database for emission lines. IX. Ionization rates, recombination rates, ionization equilibria for the elements hydrogen through zinc and updated atomic data. Astron. Astrophys 2009; 498:915–29.

[7] Jevremovic D., Dimitrijevic M.S., L.C. Popovic, Dacic M., Protic Benisek V., Bon E., et al. The project of Serbian Virtual Observatory and data for stellar atmosphere modeling. New Astron. Rev. 2009; 53: 222–6. http://stark-b.obspm.fr.

[8] Faenov A.Y., Magunov A.I., Pikuz T.A., Skobelev I.Y., Loboda P.A.. Bakshayev N.N., et al. Spectr-W-3 online database on atomic properties of atoms and ions. AIP Conf. Proc. 2002; 636: 253–62. http://spectr-w3.snz.ru.

[9] Cunto W., Mendoza C., Ochsenbein F., Zeippen C. TOPbase at the CDS. Astron. Astrophys. 1993; 275:L5–8. http://cdsweb. u-strasbg.fr/topbase/home.html.

[10] Mendoza C., Seaton M.J., Buerger P., Bellorin A., Melendez M., Gonzalez J., et al. OPserver: interactive online computations of opacities and radiative accelerations. Mon. Not. R. Astr. Soc. 2007; 378: 1031–5, http://opacities.osc.edu/.

[11] Bautista M.A., Kallman T.R. The XSTAR atomic database. Astrophys. J. Suppl. 2001; 134:139–49.

[12] RalchenkoYu., Kramida, A.E, Reader J., NIST ASD Team. NIST Atomic Spectra Database (version3.1.5), http://physics.nist.gov/asd3S. National Institute of Standards and Technology, Gaithersburg, MD; 2008.

[13] Mason N.J. Electron induced processing; applications and data needs. In: Proceedings of ICAMDATA06 AIP Conference Proceedings, vol. 901; 2007. p. 74–84.

[14] Perevalov V.I., Tashkun S.A. CDSD-296 (Carbon Dioxide Spectroscopic Databank): updated and enlarged version for atmospheric applications. In: 10th HITRAN database conference, Cambridge, MA, USA; 2008.

[15] Mikhailenko S., Barbe A., Babikov Y., Tyuterev V.G. S&MPO— a databank and information system for ozone spectroscopy on the WEB. http://smpo.iao.ru.

[16] 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.

[17] N. Jacquinet-Husson, N.A. Scott, A. Chédin, L. Crépeau, R. Armante, V. Capelle, J. Orphal, A. Coustenis, C. Boonne, N. Poulet-Crovisier, A. Barbe, M. Birk, L.R. Brown, C. Camy-Peyret, C. Claveau, K. Chance, N. Christidis, et al., The GEISA spectroscopic database: Current and future archive for Earth and planetary atmosphere studies, J. Quant. Spectr. Rad. Transfer, 2008, V. 109, Issue 6, P. 1043-1059

[18] L.S. Rothman, I.E. Gordon, R.J. Barber, H. Dothe, R.R. Gamache, A. Goldman, V.I. Perevalov, S.A. Tashkun, J. Tennyson, HITEMP, the high-temperature molecular spectroscopic database, J. Quant. Spectr. Rad. Transfer, 2010, Volume 111, Issue 15, P. 2139-2150

[19] Tennyson J., Bernath P.F., Brown L.R., Campargue A., Carleer M.R., Csaszar A.G., et al. IUPAC critical evaluation of the rotational- vibrational spectra of water vapor. PartI. Energy levels and transition Wavenumbers for $H_2{}^{17}O$, $H_2{}^{18}O$, and $HD^{16}O$. J. Quant. Spectr. Rad. Transfer, 2009;110:573–96.

[20] Malloci G., Joblin C., Mulas G. On-line database of the spectral properties of polycyclic aromatic hydrocarbons. Chem. Phys., 2007; 332:353–9. http://astrochemisty.ca.astro.it/database/

[21] Dubernet M.L., Grosjean A., Daniel F., Flower D., Roueff E., Daniel F., et al. Ro-vibrational collisional excitation database BASECOL—http://basecol.obspm.fr/ J. Plasma Fusion Res. Ser. 2006; 7: 356–57. http://basecol.obspm.fr.

[22] Schmitt B.P., Volcke E., Quirico O., Brissaud N., Fray W., Grundy J.-M., et al. GhoSST: the Grenoble astrophysics and planetology solid spectroscopy and thermodynamics database service: ‘‘RELEVANT Database’’ 2009; http://ghosst.obs.ujf-grenoble.fr/.

[23] Strazzulla G., Leto G., Palumbo M.E. Ion irradiation experiments. Adv. Space. Res. 1993; 13: 189–98 http://web.ct.astro.it/weblab/dbindex.html.

[24] Woodall J., Agundez M., Markwick-Kemper A.J., Millar T.J. The UMIST database for astrochemistry 2006. Astron. Astrophys. 2007; 466 : 1197–2003. http:// www.udfa.net.

[25] Kinetic Database for Astrochemistry, http://kida.obs.u-bordeaux1.fr.

[26] Muller H.S.P., Schloder F., Stutzki J., Winnewisser G. The cologne database for molecular spectroscopy, CDMS: a useful tool for astronomers and spectroscopists. J.Mol.Struct. 2005; 742: 215–27. http://www.ph1.uni-koeln.de/vorhersa- gen/ .

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

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

[29] А.Б. Петровский, Пространства множеств и мультимножеств, Москва, УРСС, 2003, 246С.

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

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

[32] 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.

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

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

[35] А.Ю. Ахлёстин, Н.А. Лаврентьев, М.М. Макогон, А.И. Привезенцев, А.З.Фазлиев, Инструментарий публикации данных и метаданных для распределенной информационной системы по количественной спектроскопии, Труды 11-ой Всеросс. научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL’2010, 13-17 октября 2010, Казань, 2010, С.53-59.

[36] 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.

[37] J. Tennyson, P. F. Bernath, L.R. Brown, A.Campargue, A.G. Császár, L. Daumont, at 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, V. 111, Issue 15, P. 2160-2184.

[38] A. Privezentsev, A. Fazliev, D.Tsarkov, J.Tennyson, Computed Knowledge Base for Description of Information Resources of Water Spectroscopy, Proceedings of the 7-th International Workshop on OWL: Experiences and Directions (OWLED 2010), San Francisco, California, USA, June 21-22, 2010, CEUR-WS Proc. Vol-614, Editor(s) Evren Sirin, Kendall Clark, San Francisco, 2010, http://ceur-ws.org/Vol-614/ owled2010_submission_6.pdf..