

Базы знаний для описания информационных ресурсов в молекулярной спектроскопии. 3. Формирование базовой и прикладной онтологии

Привезенцев А.И., Царьков Д.В., Фазлиев А.З.

Аннотация

Статья посвящена описанию свойств решений задач в предметной области «Количественная спектроскопия» и детализирует результаты работы [1]. Для представления свойств решений задач используется язык OWL DL. В качестве примера рассмотрена прикладная онтология источников информации о свойствах решений прямой и обратной задачи нахождения вакуумных волновых чисел в молекулярной спектроскопии диоксида углерода и его изотопологов. Показано, что создание источника информации, характеризующих свойства решения задачи, позволяет классифицировать решения задач количественной спектроскопии и организовывать семантический поиск достоверных данных в публикациях. Дано краткое описание особенностей табличного метода используемого в машине вывода FACT++.

Ключевые слова: количественная спектроскопия, язык OWL DL, вакуумные волновые числа, молекулярная спектроскопия диоксида углерода, семантический поиск.

Введение.

В библиотеках систематизация ресурсов библиотечного фонда основана на библиографических записях, определяемых тем или иным стандартом [2,3]. В цифровых библиотеках научные публикации в журналах, трудах конференций или сборниках статей являются единицами хранения и, как правило, в них содержатся системы поиска статей, основанные на библиографических записях. В работе [4, 5] описаны функциональные требования к библиотечным записям.

В научных цифровых библиотеках системы поиска опираются на тексты статей на естественном языке. Большая часть поисковых систем не использует формализованные понятия предметных областей, к которым относятся искомые ресурсы, по простой причине: большинство понятий не формализовано. Задачу поиска с учетом терминологии предметной области решают с помощью информационно-поисковых тезаурусов (см., например, [6]). В таком подходе терминология предметной области формируется с учетом лингвистических особенностей языка и онтологических отношений терминов предметной области. Существенной сложностью в формировании терминологии является ее изменение во времени. Как правило, такими изменениями пренебрегают и ограничиваются только онтологическими отношениями верхнего уровня или, если требуется детализация, онтологиями предметной области. Прикладные онтологии [7] характеризуют наиболее динамичную часть знаний и требуют для их представления описания с большим количеством деталей. В этой статье рассматривается онтология источников информации, связанных с публикациями в количественной спектроскопии. В этой предметной области причиной динамичного изменения терминологии является прогресс в измерительной аппаратуре, инициирующий создание все более сложных математических моделей молекул для изучения новых диапазонов волновых чисел.

При выполнении грантов и проектов авторы собрали опубликованные решения задач в количественной спектроскопии для ряда молекул (вода [8], диоксид углерода [9], аммиак [10], метан [11], сероводород [12] и т.д.). Части публикаций, содержащие решения одной из шести задач количественной спектроскопии, были выделены и загружены в информационные системы, каждая из которых включала в себя молекулы определенной симметрии. Созданные системы содержат все опубликованные решения задач в рамках моделей данных, описанных ранее [1]. Каждое решение является частью публикации, а в количественной спектроскопии основной частью, так как содержит наибольшее число типизованных фактов. В простом случае решение задачи можно принять за модель публикации. База данных по молекуле диоксида углерода, описанная в предыдущей части статьи, и содержащая опубликованные решения задач, может быть рассмотрена как модель электронной библиотеки. Решение задачи, дополненное свойствами, может служить более точной формальной моделью публикации. В спектроскопии такой набор свойств был предложен в [13].

В этой статье детально рассмотрена модель публикации для предметной области «Количественная спектроскопия». Она основана на двух понятиях: «первичный источник данных» и «источник информации». Источник информации содержит в себе ссылку на один или несколько источников данных, являющихся частями публикации. Источник информации может включать в себя высказывания, не содержащиеся в публикациях, к которым относятся источники данных, связанные с ним. Определения понятий представлены в первом параграфе статьи. Во втором параграфе в рамках языка OWL DL описаны классы, свойства и индивиды базовой онтологии для количественной спектроскопии и связанных с ней двух предметных областей. В третьем параграфе описаны классы и свойства прикладной онтологии, относящейся к двум задачам количественной спектроскопии. В четвертом параграфе в разных представлениях представлен элемент класса, характеризующий свойства решения задачи T6. Эти свойства могут быть как индивидуальными свойствами (относящимися к источнику данных из одной публикации), так и относительными свойствами этого источника по отношению к источникам данных из иных публикаций. Отмечено сходство и различие введенного в работе понятия «источник информации» с библиографической записью в цифровых библиотеках. В последнем параграфе приведено краткое описание табличного метода используемого в машине вывода FACT++. Она используется при работе со всеми онтологиями ИВС.

1. Предметная область «Источник информации»

Основная информация поступает к исследователям из опубликованных статей, написанных на естественном языке. Технической задачей при работе со статьями является построение субъектно-предикатных структур [14]. В таких структурах элементарным информационным объектом является высказывание. Представляемое в языке OWL DL высказывание состоит из трех частей: субъект, свойство, объект. Если в высказываниях, относящихся к некоторой предметной области, объект может быть также субъектом (обладает свойствами, относящимися к этой предметной области), то свойство, значением которого является этот объект, называют объектным [15]. Если же

объект не обладает никакими свойствами, относящимися к предметной области, и определяется одним из стандартных типов данных, то такое свойство называется конкретным свойством (Datatype). Именно эти свойства характерны для количественных наук, в которых подобные высказывания составляют подавляющее большинство. Примером высказывания, содержащего конкретное свойство, является высказывание «частота имеет значение (равна) 17492.01». В публикациях, относящихся к количественным наукам, такого рода высказывания составляют большую часть текста. На рис.1 показан график зависимости количества высказываний от субъектов высказываний, относящихся к предметной области, для работы [16].

В нулевом приближении при представлении публикации в информационной системе имеет смысл ограничиться высказываниями с теми физическими сущностями, которые являются субъектами высказываний в большем числе высказываний в публикации. В количественной спектроскопии число этих сущностей достаточно велико, почти все они описаны в модели данных количественной спектроскопии [1].

Можно ввести количественную характеристику, относящуюся к публикации

$$F(A_i) = (Q(A_i) / (\sum_{i=1}^n Q(A_i))) / (\max (Q(A_i) / (\sum_{i=1}^n Q(A_i))))$$

где A_i – термин или понятие, встречающиеся в высказываниях статьи в качестве субъекта, $Q(A)$ – число высказываний, в которых A является субъектом, n – число терминов и понятий. Схема источников данных, извлеченных из статей, содержит понятия и термины для которых $F(A_i)$ близко к единице.



Рис. 1. Зависимость числа высказываний от субъектов количественной спектроскопии

Разнообразие молекул, для которых решались задачи, выделенные в работе [1], и методов, которыми они решались, достаточно большое. По этой причине в одной публикации могут быть приведены решения нескольких задач разными методами и для разных молекул или их изотопологов. При систематизации данных, извлеченных из публикаций, такое смешение создает много проблем. По этой причине в работе используется информационный объект, представляющий оригинальные данные публикации, относящиеся к одной молекуле, одной задаче спектроскопии и одному методу решения.

Определение 1. Все части опубликованного решения задачи количественной спектроскопии, дополненные названием молекулы, библиографической ссылкой и названием метода решения задачи (или ссылкой на описание метода) называются *первичным источником данных*.

Мы предполагаем, что пустые решения не публикуются. С другой стороны решения задач могут содержать данные измерений, которые со временем устаревают. Источник данных, содержание которого целиком отклонено экспертами будем называть ничтожным. Количество таких источников в современной спектроскопии незначительно.

Формализованный первичный источник данных содержит решение задачи и обладает свойствами (*isSolutionOf*, *hasMethod*, *isRelatedToSubstance* и *hasReference*) имеющими кардинальность равную 1. Важной характеристикой источника данных является независимость значений этих свойств от времени. Ключевым свойством в определении источника данных является *hasReference*. Оно тесно связано с определением информационного ресурса [17, 18], и значения свойства *hasReference* образуют подкласс класса «Информационный ресурс».

В количественной спектроскопии, наряду с журналами, монографиями, отчетами и трудами конференций, в последнее десятилетие появились публикации решений задач в Вебе. Необходимость публикации в Вебе обусловлена значительными их объемами (превышающими сотни Гб.) Примерами таких ресурсов являются спектральные данные, размещенные в Европе [19,20], России [21], США [22,23] и т.д.

Первичные источники данных, относящиеся к одной публикации, не имеют общих данных. Этот факт схематично представлен на рис.2а, где овалом обозначена публикация, а треугольниками - источники данных. В публикации по количественной спектроскопии может содержаться не один первичный источник данных.

Определение 2. Информационный объект, обладающий базовыми свойствами первичного источника данных, кардинальность любого из которых отличается от единицы, называется *составным источником данных*.

Примером составного источника данных является любой экспертный массив спектральных данных (например, Hitran [23]).

Первичный источник можно наделять дополнительными свойствами. Перечень и число этих свойств зависит от информационных задач, для решения которые используются такие свойства. Источник данных с дополнительными свойствами назовем источником информации.

Определение 3. Первичный источник данных, наделенный дополнительными свойствами, называется *первичным источником информации*, извлеченной из публикации.

Источник информации представляет собой набор свойств и их значений, относящихся к источнику данных. Для ряда информационных задач, например, задачи поиска достоверных решений задач количественной спектроскопии, можно выбрать свойства, значения которых вычисляются автоматически. Как правило, источник информации включает в себя некоторые высказывания из публикации, содержащей источник данных, который он описывает. Большая часть источника информации характеризует ту часть знания, которая содержится в публикации в неявном виде.

Перечень дополнительных свойств определяется исследователем исходя из информационных задач, которые ему необходимо решать. В нашей работе таких задач две. Это задача семантического поиска и задача автоматического построения экспертного массива данных. Заметим, что первичные источники информации, относящиеся к одной публикации, не содержат идентичных высказываний. Различие между публикацией и первичным источником информации может быть существенно меньшим по сравнению с различием между публикацией и первичным источником данных. Различие обусловлено теми дополнительными свойствами решения задачи в публикации, которые вошли в определение того или иного источника информации. Например, такими дополнительными свойствами могут быть описание достоверности решения задачи, описание стандартных отклонений исходного источника данных от других источников данных и т.д.. Кроме того, высказывания, содержащиеся в первичном источнике информации, могут не содержаться в публикации. Это утверждение демонстрируется на рис.2b, на котором шестиугольником обозначен источник информации.

Для решения задач поиска и построения экспертного массива выбраны несколько групп свойств. Классифицируем эти группы в соответствии с онтологией верхнего уровня [24]. К группе индивидуальных (независимых) свойств отнесем свойства, значения которых не зависят от иных источников данных. К группе относительных свойств отнесем свойства, характеризующие пару источников данных. К группе континуантов отнесем свойства, значения которых не зависят от времени, а к группе оккurrentов отнесем свойства, значения которых зависят от времени. Следуя работе [24] независимые свойства, значения которых не меняются со временем, назовем схемами, а зависящие – скриптами. Независимые от времени относительные свойства назовем описаниями, а зависящие – историями.

К схемам относятся свойства характеризующие:

1. физические величины (число уровней энергии, число вакуумных волновых чисел, и т.д);
2. интервалы изменения физических величин;
3. число переходов без квантовых чисел, число дубликатов, число переходов, не удовлетворяющих правилам отбора и т.д.
4. тип источников данных и информации.

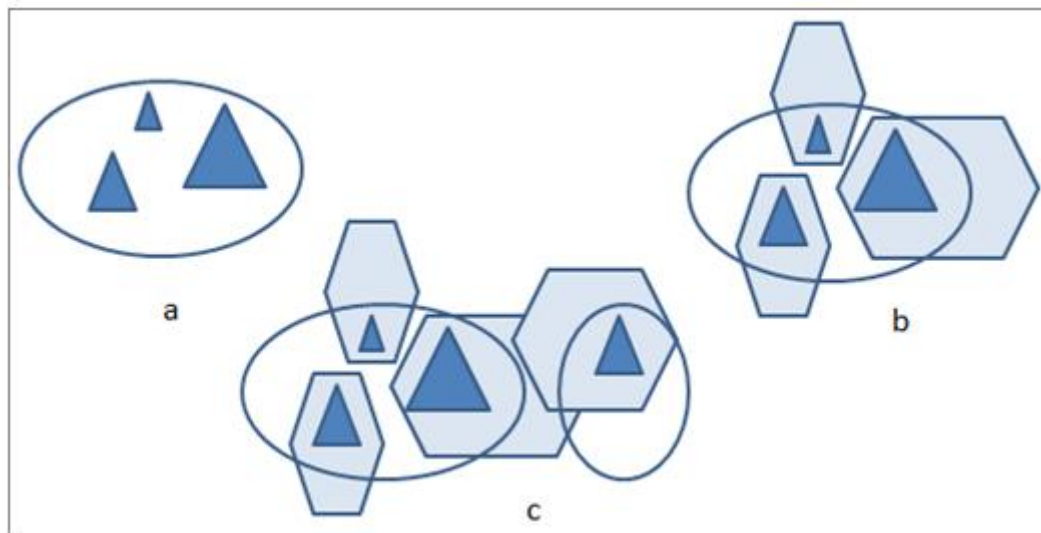
К скриптам относится такое свойство как число признанных экспертами не достоверных переходов или состояний.

К описаниям относятся максимальное отклонение волновых чисел идентичных переходов в сравниваемой паре первичных источников данных, число идентичных колебательных полос, содержащихся в паре сравниваемых первичных источников данных, число переходов, для которых нарушен порядок следования и т.д.

К историям относятся свойства предыдущей группы в случае, когда сравниваются достоверные части источников данных с учетом отклоненных экспертами данных.

Источник информации является сложным информационным объектом. Он может содержать, как описания, так и истории. Наличие той или иной части обусловлено информационными задачами для решения которых он создается. В данной работе изучаются конкретные источники информации, всегда связанные с источником данных (другими словами кардинальность свойства *hasDataSource*, доменом которого является класс источников информации, равен 1).

На рис.2 показан случай, когда источник информации содержит высказывания из другой публикации.



2. Базовая онтология источников информации по спектроскопии молекулы диоксида углерода

В статье [1] описаны данные, представляющие в выбранной модели количественной спектроскопии решения шести задач. Решения задач обладают свойствами. Эти свойства и их значения связаны с объектами и схемами молекулярной спектроскопии, например: физические величины, определенные в решениях; математические описания, характеризующие их значения; публикация этих решений. Выбор набора свойств определяется информационными задачами, для решения которых они необходимы.

Основными информационными задачами, в которых применяются свойства данных, являются задачи поиска и интеграции, как источников данных, так и источников информации.

Рассматривая свойства решений задач количественной спектроскопии, как основной ресурс для систематизации, мы опираемся на онтологию задач [25, 26]. В онтологии задач верхнего уровня [26] основными концептами являются цель решения задачи, предметные данные, описывающие экземпляр решения задачи, состояние решения задачи и знание о решении задачи. В этой главе основное внимание сосредоточено на создании прикладной онтологии источников информации о решении задач количественной спектроскопии.

В ИВС по количественной спектроскопии создано несколько версий прикладных онтологий по количественной спектроскопии. Они отличаются как наборами индивидов, так и таксономиями классов. Ниже описана последняя на данный момент версия №6. Названия всех классов начинаются с заглавной буквы. Названия всех свойств начинаются с маленькой буквы (в подавляющем большинстве они начинаются с has или is). Для их представления используется наклонный шрифт. Объектные свойства выделены жирным шрифтом.

2.1. Свойства и классы базовой онтологии источников информации количественной спектроскопии

За основу для построения онтологии задач принят подход, в котором задача является системой, для описания которой используется IPO (Input-Processing-Output) - модель. Так, к числу включенных в рассмотрение метаданных для входных и выходных данных относятся их интенционалы и ряд атрибутов, характеризующих количественную сторону экстенционала данных. В выбранной модели метаданные для входных данных представляют собой ссылки на URI ресурсы. Количественные значения, содержащиеся в метаданных, формируются динамически при загрузке файлов пользователя только для выходных данных задачи.

2.1.1. Классы базовой онтологии

При формировании класса в OWL DL [27] класс можно отнести к одному из 6 типов (D1 - задание класса идентификатором, D2 - задание класса перечислением всех элементов, D3 - задание класса ограничением на свойства, D4 - задание класса объединением классов, D5 - задание класса пересечением, D6 - задание класса дополнением).

Таблица 1.

Класс	Суперкласс	Способ определения	Предметная область
Method	Thing	D1	General
Task	Thing	D2	
InputData	Thing	D1	
OutputData=ResultOfTaskSolution	Thing	D1	
State	Thing	D2	
Process	Thing	D2	
Substance	Thing	D1	Вещество
SymmetryGroup	Thing	D2	
PhysicalState	Thing	D2	
PhysicalQuantity	Thing	D1	Количественная спектроскопия
DimensionalQuantity	PhysicalQuantity	D1	
Unit	PhysicalQuantity	D1	
QuantumNumbers	PhysicalQuantity	D1	
QuantumNumbersType	PhysicalQuantity	D1	

DataSource	Thing	D3	Информационные объекты
Metadata=DataProperties	Thing	D1	
InputData_MD	Metadata	D3	
OutputData_MD	Metadata	D3	
InformationSource	DataSource	D3	
DataSourcesPair	OutputData_MD	D1	
OrderingPair	DataSourcesPair	D1	
SDPair	DataSourcesPair	D1	

В Таб.1. жирным шрифтом выделены классы по описанию типа D2. Они содержат только перечисленные элементы. Другими словами, в онтологии рассматривается ограниченный круг задач количественной спектроскопии, каждая из которых характеризуется экземпляром класса **Task**. Решение каждой из таких задач характеризует физические характеристики, свойственные определенному физическому процессу или состоянию. В онтологии содержатся источники информации только о тех веществах, группа симметрии которых указана в классе **SymmetryGroup**. Число представлений квантовых чисел, используемых в источниках информации, конечно и исчерпывается в онтологии элементами класса **QuantumNumbersType**.

В этой онтологии можно выделить три группы классов:

Классы, содержащие объекты, относящиеся к предметной области «Вещество», в частности, молекулы воды и ее изотопологи (**Substance**, **PhysicalState**, **SymmetryGroup**).

Классы, содержащие математические модели объектов, используемые в молекулярной спектроскопии (**SymmetryGroup**, **Task**, **Method**, **InputData**, **OutputData**).

Классы, содержащие объекты, относящиеся к предметно области «Количественная спектроскопия», (**PhysicalQuantity**, **DimensionalQuantity**, **QuantumNumbersType**, **Unit**, **QuantumNumbers**)

Классы, содержащие информационные объекты (**Metadata**, **InformationSource**, **DataSource**, **DataSourcesPair**, ...).

Приведем описание классов по группам.

1. Классы предметной области «Вещество».

Класс **Substance** содержит сущность «воздух» и химические элементы, набор молекул и их изотопомеров, характерных для атмосфер планет. Класс **PhysicalState** состоит из перечисления физических состояний вещества в исследуемой среде (изолированный атом и молекула, газовое состояние). Класс **SymmetryGroup** содержит элементы, соответствующие точечным группам, используемым при описании молекул. Класс **DataSourcesPair** содержит индивиды, характеризующие отношения пары источников данных.

2. Базовые классы предметной области «Количественная спектроскопия».

Класс **PhysicalQuantity** состоит из перечисления физических величин предметной области используемых при решении задач (уровень энергии, интенсивность, вакуумная частота, и т.д.).

Класс **DimensionalQuantity** описывает сущности, которые имеют единицы измерения. Он является подклассом **PhysicalQuantity** и является доменом для свойства **hasUnit**.

Класс **QuantumNumbersType** состоит из перечисления типов квантовых чисел, используемых для идентификации уровней энергии и переходов исследуемого вещества (нормальные моды, BT2, Швенке).

Класс **Unit** состоит из перечисления единиц измерения физических величин, используемых при решении задач T1-T7.

Класс **QuantumNumbers** состоит из индивидов каждый из которых характеризует колебательное состояние или колебательно-вращательную полосу молекулы.

3. Описание классов математических объектов, используемых в онтологии.

Класс **Task** состоит из перечисления названий задач предметной области.

Класс **Method** состоит из перечисления методов предметной области, используемых при решении задач.

4. Классы, содержащие информационные объекты.

Класс **Metadata** является базовым для классов содержащих метаданные по всем физическим величинам, используемым в данной онтологии.

Класс **InformationSource** является ключевым и содержит индивиды, описывающие результаты решений предметных задач. Он задаётся ограничением на свойство **hasOutputData_MD** exactly 1.

Класс **DataSource** содержит элементы, содержащие в себе результаты решения задач, описание вещества, методов решения задач и библиографическое описание работы, в которой опубликованы эти результаты.

Класс **OrderingPairRelation** содержит описания относительных некорректностей, связанных с нарушением упорядочения по значениям сравниваемых физических величин. Детали об экземплярах этого класса рассмотрены в пятой части статьи.

Класс **PairRelation** содержит все используемые парные описания в предметной области для которой создается прикладная онтология.

Заметим, что далее, для определенности, названия классов и свойств с окончанием **_MD**, относятся к классам, содержащим описания решений задач, или к свойствам, значения которых являются описаниями решений задач.

2.1.2. Свойства базовой онтологии

В OWL DL-онтологии определяются две группы свойств: объектные свойства и свойства типов данных.

Таблица 2. Объектные свойства базовой онтологии и их области определения и значений.

ObjectProperty	Domain	Range
hasMethod	Task	Method
hasInputData	Task	InputData
hasInputData	Task	OutputData=SolutionOfTask
isSolutionOf	DataSource or OutputData	Task
hasSubstance	DataSource or InputData or OutputData	Substance
hasSymmetryGroup	Substance	SymmetryGroup
hasPhysicalState	Substance	PhysicalState
isInteractingSubstance	Substance	Substance
isPhysicalState	PhysicalState	Substance
hasUnit	PhysicalQuantity	Unit
hasQuantumNumbersType	QuantumNumbers	QuantumNumbersType
hasDataSourceURI	InformationSource	DataSource
hasInputData_MD	InformationSource	InputData_MD
hasOutputData_MD	InformationSource	OutputData_MD

Опишем некоторые объектные свойства из Таб.2.

Объектное свойство *isSolutionOf* характеризует, к какой задаче относится источник данных или выходные данные. Объектное свойство *hasSubstance* применяется к входным и выходным данным задач класса **Task** и источникам данных. Источники информации представляют собой описания того или иного набора свойств конкретного источника данных. Свойство *hasDataSourceURI* указывает, к какому источнику данных относится это описание. Объектное свойство *hasInputData_MD* применяется к экземплярам класса *InformationSource* и его значение описывает свойства входных данных, использованных при решении задачи.

К свойствам типов данных относятся свойства, показанные в таблице 3.2.

Таблица 3. Свойства типов данных, используемые в прикладной онтологии

DatatypeProperty	Domain	Range
hasUncertainty	DimensionalQuantity	float
hasReference	DataSource	string
hasDOI	DataSource	string
hasURL	DataSource	string

Свойства типов данных имеют следующую семантику:

- *hasUncertainty* задаёт неопределенность в измерении физической величины;
- *hasReference* определяет библиографическую ссылку на публикацию, в которой опубликовано решение задачи;
- значение свойства *hasDOI* однозначно характеризует информационный ресурс;
- *hasURL* определяет адрес в вебе соответствующей публикации.

2.1.3. Индивиды базовых классов, определенных перечислением

Класс «Task» содержит индивиды, соответствующие задачам, описанным в работе [1] (T1, T2, T3, T4, T6, T5, T7, ET1).

Класс «PhysicalState» содержит три фазовых состояния (Gas, SingleAtom, SingleMolecule).

Классы «Process» и «State» содержат индивиды, соответствующие физическим процессам, в которых принимают участие вещества, спектральные свойства которых представляют интерес для пользователя.

Класс «QuantumNumbersType» содержит индивиды, характеризующие типы квантовых чисел (вариационные квантовые числа (BT2, VTT), нотации (HitranNotation, CDSNotation), нормальные моды (NormalModes) и т.д.).

Экземпляры этих классов задаются однократно и присутствуют во всех версиях онтологий. Экземпляры практически всех классов базовой онтологии определяются вручную. В прикладных онтологиях подавляющее большинство индивидов формируется автоматически.

3. Классы и свойства прикладной онтологии задачи T6

Описываемая в данном параграфе онтология ориентирована на решение задачи поиска достоверных источников данных количественной спектроскопии и систематизации источников данных по величине среднеквадратических отклонений в распределенной информационной системе. Здесь мы ограничились описанием свойств решений двух задач (T2 и T6) спектроскопии диоксида углерода. В задаче поиска достоверных источников данных предполагается, что достоверность решений определяется из проверки их соответствия правилам отбора, которые являются ограничениями, накладываемыми на квантовые числа.

В прикладной онтологии элементы практически всех классов создаются автоматически. Опишем некоторые из классов, которые определяются нами не в автоматическом режиме. Экземплярами классов TransitionQuantumNumbers_MD и VacuumWavenumbers_MD являются описания, соответственно, квантовых чисел и вакуумных волновых чисел относящихся к решениям задачи в источниках данных.

В прикладной онтологии информационных ресурсов, относящихся к задачам определения вакуумных волновых чисел, появляется ряд классов, связанных с понятием колебательно-вращательная полоса (ro-vibrational band), возникающим в моделях молекул, трактующих ее как набор частей совершающих колебательные и вращательные движения. Эти движения типизируются с помощью набора квантовых чисел, обязательно содержащих в себе все квантовые числа, относящиеся к описанию колебательного состояния.

Элемент класса RoVibrationalBand представляет полосу, встречающуюся в публикациях.

Элемент класса RoVibrationalBand_MD описывает свойства конкретной полосы из публикаций.

Элемент класса RoVibrationalBandQuantumNumbersList характеризует нотацию, примененную для описания полосы и набор квантовых чисел, связанный с понятием колебательно-вращательная полоса.

Класс Tn-IS содержит описания свойств всех физических величин, связанных с входными и выходными данными задачи Tn, (n = 2,6).

Таблица 4. Классы прикладной онтологии задачи T6.		
Класс	Суперкласс	Предметная область
TransitionQuantumNumbers_MD	T2-OutputData_MD or T6-OutputData_MD	Количественная спектроскопия
VacuumWavenumbers_MD	T2-OutputData_MD or T6-OutputData_MD	
EinsteinCoefficient_MD	T2-OutputData_MD or T6-OutputData_MD	
RoVibrationalBandQuantumNumbersList	PhysicalQuantity	
RoVibrationalBandQuantumNumbers	PhysicalQuantity	
RoVibrationalBand	PhysicalQuantity	
RoVibrationalBand_MD	PhysicalQuantity	
T6-IS	InformationSource	Информационные объекты
T6-InputData_MD	InputData_MD	
T6-OutputData_MD	OutputData_MD	
RoVibrational BandsPair	DataSourcePair	
SDRoVibrationalBandsPair	SDPair	

Элемент класса **SDRoVibrationalBand** описывает свойства двух идентичных полос в паре источников данных из разных публикаций, для которых вычисляется стандартное отклонение. К числу этих свойств относятся тип нотации, используемый для представления колебательно-вращательных полос, число переходов с тождественными наборами квантовых чисел в полосах, максимальное значение разности вакуумных волновых чисел идентичных переходов и величина стандартного отклонения сравниваемых полос.

Элемент класса **RoVibrational BandsPair** описывает все пары совпадающих полос, относящихся к сравниваемым источникам данных и указывает на число совпадающих полос.

Доменами всех объектных свойств этой прикладной онтологии, а также областями значений трех из них являются описания (части источников информации).

Таблица 5. Объектные свойства прикладной онтологии.

ObjectProperty	Domain	Range
<i>hasRoVibrationalBandQuantumNumbersList</i>	RoVibrationalBand_MD	RoVibrationalBandQuantumNumbersList
<i>hasRoVibrationalBand_MD</i>	TransitionsQuantumNumbers_MD	RoVibrationalBand_MD
<i>hasQuantumNumbersOfBand</i>	RoVibrationalBand_MD	RoVibrationalBand
<i>hasVacuumWavenumbers_MD</i>	OutputData_MD	VacuumWavenumbers_MD
<i>hasEinsteinCoefficient_MD</i>	OutputData_MD	EinsteinCoefficient_MD
<i>has SDRoVibrational BandsPair</i>	DataSourcesPair_MD	SDRoVibrationalBandsPair_MD
<i>isMemberOf SDRoVibrationalBandsPair</i>	SDRoVibrationalBandsPair_MD	RoVibrationalBand
<i>hasSDMember</i>	DataSourcesPair_MD	DataSource
<i>hasSDRoVibrationalBandQuantumNumberList</i>	SDRoVibrationalBandsPair_MD	RoVibrationalBandQuantumNumbersList

Конкретные свойства (Datatype) онтологии приведены в таблицах 6 и 7. В таб.5 содержатся свойства, которые не зависят от конкретной молекулы, для которой была решена задача. Перечень этих свойств неполный. Поясним некоторые свойства, относящиеся к квантовым числам. Они указывают на число достоверных (valid), уникальных (unique), или не полностью определенных (unassigned) переходов. Эти числа вычисляются по известным алгоритмам. Значение свойства *hasNumberOfRejectedTransitions* не вычисляется и может изменяться со временем. Это связано со следующими причинами. С одной стороны, квантовые числа в решении могут удовлетворять правилам отбора, но быть неправильно приписанными к переходу или состоянию. Эксперты указывают на неправильное отнесение. С другой стороны, по мере совершенства вычислений и способов измерений значение вакуумного волнового числа, интенсивности или других характеристик может не соответствовать точности требуемой для решения прикладных задач.

Таблица 6. Конкретные индивидуальные свойства, не зависящие от типа молекулы.

DatatypeProperty	Domain	Range
<i>hasNumberOfRoVibrationalBands</i>	RoVibrationalBand_MD	integer
<i>hasEinsteinCoefficient</i>	EinsteinCoefficient_MD	boolean
<i>hasMaxVacuumWavenumberValue</i>	VacuumWavenumbers_MD	float
<i>hasMinVacuumWavenumberValue</i>	VacuumWavenumbers_MD	float
<i>hasNumberOfVacuumWavenumbers</i>	VacuumWavenumbers_MD	integer
<i>hasNumberOfAllValidIdentifications</i>	TransitionQuantumNumbers_MD	integer
<i>hasNumberOfAllInvalidIdentifications</i>	TransitionQuantumNumbers_MD	integer
<i>hasNumberOfNonuniqueTransitions</i>	TransitionQuantumNumbers_MD	integer
<i>hasNumberOfUniqueTransitions</i>	TransitionQuantumNumbers_MD	integer
<i>hasNumberOfDisorderedTansitions</i>	DataSourcePair	integer
<i>hasMaxDifferenceValueOfVacuumWavenumbersBandsPair</i>	DataSourcePair	float
<i>hasTotalNumberOfIdenticalTransitions</i>	DataSourcePair	integer
<i>hasTotalSDValue</i>	DataSourcePair	float
<i>hasSDValue</i>	SDRoVibrationalBandPair	float

Наконец, необходимо ввести значительное число конкретных свойств, вычисление значений которых зависит от исследуемой молекулы, нотации квантовых чисел, используемой при ее описании, и физического процесса, спектральные свойства которого изучаются. В спектроскопии значения всех этих свойств вычисляются при проверке правил отбора для каждого перехода. Правила отбора являются следствиями математической модели молекулы, ориентированной на описание того или иного физического процесса. Поскольку молекул, исследуемых в количественной спектроскопии, существенно больше сотни, а нотация квантовых чисел, как правило, не одна, то количество свойств, относящихся к описанию числа переходов, является достаточно большим. Для типизации имен этих свойств выбрана следующая структура имени свойства

hasNumberOfInvalid(Valid)Transitions-XYZN.

Здесь X - принимает два значения (Т,S), которые характеризуют переход и состояние, Y - принимает значения, связанные с индивидами класса Process (Ab-поглощение, Em-испускание, Ra-рамановское излучение и т.д.), Z - обозначает нотацию, использованную при описании молекулы (см. Таб.8), и N - номер правила отбора в реестре правил отбора (см. Таб.9).

Приведем пример. При вычислении значения свойства *hasNumberOfInvalidTransitions-TAbNM3* каждый переход проверяется на соответствие правилу отбора *TAbNM3*, то есть, если значение полного углового момента не изменилось в результате перехода, а

симметрия ЭПСИЛОН изменилась, то правило отбора нарушается, и значение этого свойства увеличивается на 1.

Таблица 7. Конкретные индивидуальные свойства, зависящие от вещества, нотации для его описания и физических процессов (на примере молекулы диоксида углерода).

DatatypeProperty	Domain	Range
hasNumberOfInvalidTransitions-TAbNM_1	TransitionQuantumNumbers_MD	integer
hasNumberOfInvalidTransitions-TAbNM_2	TransitionQuantumNumbers_MD	integer
hasNumberOfInvalidTransitions-TAbNM_3	TransitionQuantumNumbers_MD	integer
hasNumberOfInvalidTransitions-TAbNM_4	TransitionQuantumNumbers_MD	integer
hasNumberOfInvalidTransitions-TAbNM_5	TransitionQuantumNumbers_MD	integer
hasNumberOfInvalidTransitions-TAbNM_6	TransitionQuantumNumbers_MD	integer

Таблица 8. Часть реестра нотаций для представления квантовых чисел переходов.

Нотации	Субъект	Молекулы	Квантовые числа
Нотации CDSД (NC)	T	$^{12}\text{C}^{16}\text{O}_2, ^{12}\text{C}^{17}\text{O}_2, ^{12}\text{C}^{18}\text{O}_2, ^{13}\text{C}^{16}\text{O}_2, ^{13}\text{C}^{17}\text{O}_2, ^{13}\text{C}^{18}\text{O}_2$	P', n' J', ε' - P'', n'' J'', ε''
Нотации CDSД (NCB)	T	$^{12}\text{C}^{16}\text{O}_2, ^{12}\text{C}^{17}\text{O}_2, ^{12}\text{C}^{18}\text{O}_2, ^{13}\text{C}^{16}\text{O}_2, ^{13}\text{C}^{17}\text{O}_2, ^{13}\text{C}^{18}\text{O}_2$	P', n' Br - P'', n'' J'', ε" Значения принимаемые Br Q : Q = J' - J'' = 0 = ΔJ P : P = J' - J'' = -1 = ΔJ R : R = J' - J'' = +1 = ΔJ
Нотации Hitran (NH)	T	$^{12}\text{C}^{16}\text{O}_2, ^{12}\text{C}^{17}\text{O}_2, ^{12}\text{C}^{18}\text{O}_2, ^{13}\text{C}^{16}\text{O}_2, ^{13}\text{C}^{17}\text{O}_2, ^{13}\text{C}^{18}\text{O}_2$	(верхнее (') и нижнее (") состояния) $\nu_1', \nu_2', l_2', \nu_3', r' J', \epsilon' - \nu_1'', \nu_2'', l_2'', \nu_3'', r'' J'', \epsilon''$ Колебательные квантовые числа ν_1 - кратность возбуждения 1-го колебания, целое. ($\nu_{1\min}=0, \nu_{1\max}=99$) ν_2 - кратность возбуждения 2-го колебания, целое. ($\nu_{2\min}=0, \nu_{2\max}=99$) ν_3 - кратность возбуждения 3-го колебания, целое. ($\nu_{3\min}=0, \nu_{3\max}=99$) l_2 - колебательный угловой момент ($l_{2\min}=0, l_{2\max}=99$) r - номер компоненты Ферми-мультиплета, целое. ($r_{\min}=0, r_{\max}=99$) Вращательные квантовые числа J - полный угловой момент ($J_{\min}=0, J_{\max}=999$) ε - симметрия (e, f)
Нотации Hitran (NHB)	T	$^{12}\text{C}^{16}\text{O}_2, ^{12}\text{C}^{17}\text{O}_2, ^{12}\text{C}^{18}\text{O}_2, ^{13}\text{C}^{16}\text{O}_2, ^{13}\text{C}^{17}\text{O}_2, ^{13}\text{C}^{18}\text{O}_2$	(верхнее (') и нижнее (") состояния) $\nu_1', \nu_2', l_2', \nu_3', r' \text{Br} - \nu_1'', \nu_2'', l_2'', \nu_3'', r'' J'', \epsilon''$

Таблица 9. Реестр правил отбора для переходов в нотации Нотации Hitran (Br) для молекулы диоксид углерода.

Субъект	Ab	NM	Номер правила	Правило	Молекулы
T	Ab	NM	1	Разрешенные переходы $\Delta J = 0$ или -1 или +1	$^{12}\text{C}^{16}\text{O}_2$, все изотопологи
T	Ab	NM	2	Разрешенные переходы Если $\Delta J = 0$, то симметрия должна измениться (e <-- f, f <-- e)	$^{12}\text{C}^{16}\text{O}_2$, все изотопологи
T	Ab	NM	3	Разрешенные переходы Если $\Delta J \neq 0$, то симметрия не изменяется (e <-- e f <-- f)	$^{12}\text{C}^{16}\text{O}_2$, все изотопологи
T	Ab	NM	4	Запрещенные переходы если $l_2' = l_2'' = 0$, то независимо от J - Q - ветвь запрещена.	$^{12}\text{C}^{16}\text{O}_2$, все изотопологи
T	Ab	NM	5	$\Delta \nu_1 = \nu_1' - \nu_1'' , \Delta \nu_2 + \Delta \nu_3 = 2n+1, n=0,1,2,3, \dots$	$^{12}\text{C}^{16}\text{O}_2, ^{12}\text{C}^{17}\text{O}_2, ^{12}\text{C}^{18}\text{O}_2, ^{13}\text{C}^{16}\text{O}_2, ^{13}\text{C}^{17}\text{O}_2, ^{13}\text{C}^{18}\text{O}_2$

Стандартные отклонения	
Тип: HitranNotation	Число источников данных [17]
Источник данных	Частота
1983_EsRo_C_1802	$a = 7.500e-3$ $d = [0]$ $b = 1.702e-3$ $c = [62]$ $e = [2]$ $f = [1]$
1971_ObRaJoGo_C_1802	$8.236e+1$ [18] $1.833e+1$ [664] [19] [12]

Рис.4. Часть визуализированных метаданных, относящихся к вычислению среднеквадратических значений

Здесь a - максимальная разность значений вакуумных волновых чисел идентичных переходов, b - суммарное стандартное отклонение по всем полосам, c - число вакуумных волновых чисел с совпавшими квантовыми числами, d - число несовпадений при сравнении порядка следования вакуумных частот, e (f) - число полос типа 1(2), совпавших в источниках данных.

Подобного рода визуализация ориентирована на интерпретацию свойств исследователем. При автоматизированной обработке свойств решений интерес представляют типы субъектно-предикатных структур, соответствующих элементам класса T6-IS. Эти структуры изменяются в зависимости от молекулы, для которой решается задача, выбранных нотаций и физических процессов изучаемых в количественной спектроскопии. Представление индивидов классов в виде кода OWL-онтологии можно найти на сайте SP@DIS [29]. Выделим две группы свойств. Значения индивидуальных свойств элемента класса T6-IS не зависят от наличия в этом классе каких-либо иных элементов. Значения относительных свойств зависят от некоторых элементов того же класса.

Отметим правила формирования имен индивидов, создаваемых в ИС автоматически. Рассмотрим имя индивида **V5_T6_71_2007_ToMiBrDe_C_1802**. Оно содержит ряд условных обозначений: **V5** - номер версии онтологии, **T6** - задача, решение которой содержится в описываемом источнике данных, **71** - номер источника данных в базе данных, **2007** - год публикации решения, **ToMiBrDe** - аббревиатура фамилий первых 4 авторов статьи и **C_1802** - символы, характеризующие исследованную молекулу, в нашем случае изотополог диоксида углерода C1802.

4.1. Индивидуальные свойства элемента класса T6-IS

На рис.4 показана часть элемента класса T6-IS содержащая только индивидуальные свойства.

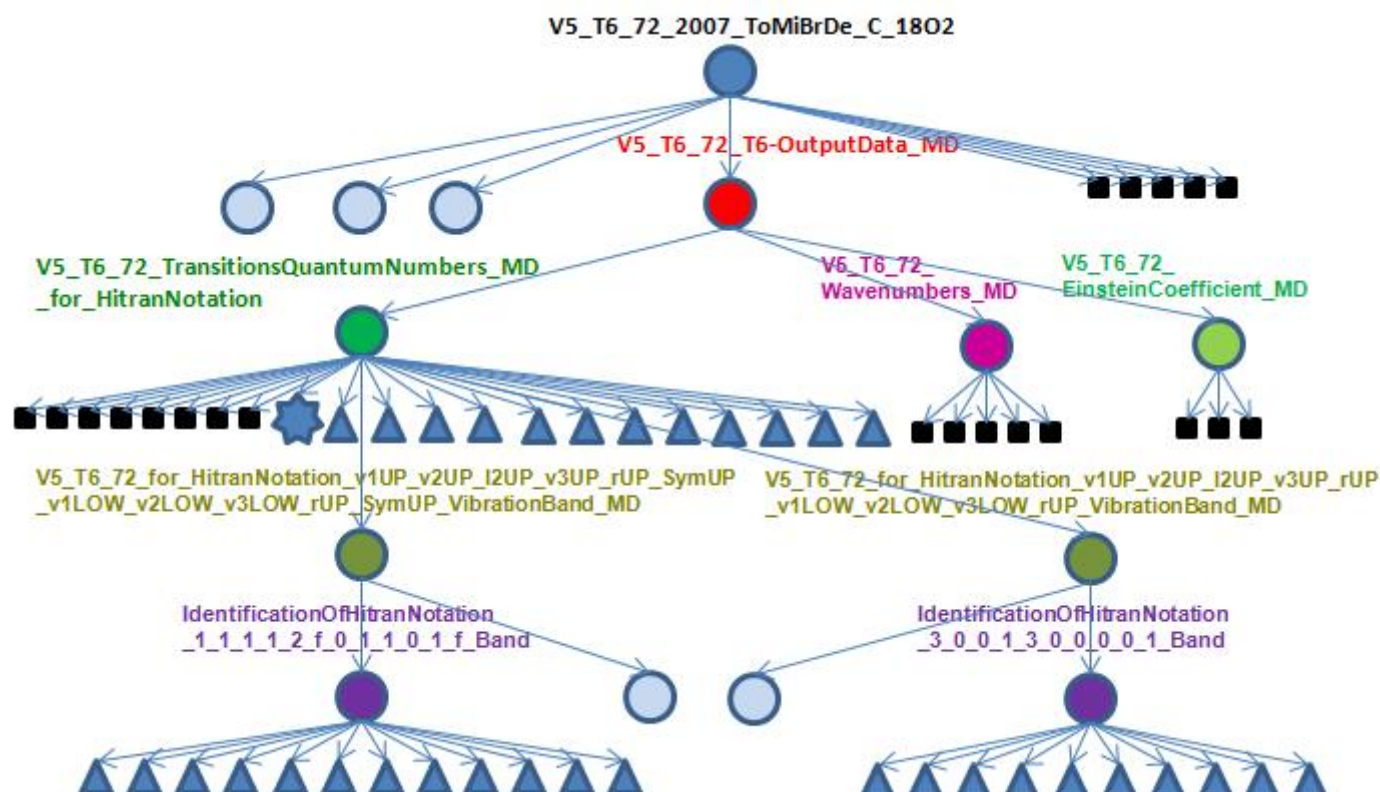


Рис.5. Субъектно-предикатная структура, включающая все индивидуальные свойства элемента класса T6-IS.

На рис.5 показана субъектно-предикатная структура, соответствующая части элемента, визуализированного на рис.3. Она состоит из

высказываний и содержит только индивидуальные свойства источника данных. Стрелка соответствует свойству. Окружностями обозначены индивиды онтологии, а треугольники, квадраты и восьмиугольники являются значениями конкретных свойств. Семиугольником выделено свойство (число переходов в источнике информации, отклоненных экспертами), значение которого может меняться со временем.

Треугольники описывают свойства, характерные только для исследуемой молекулы, применяемой к ней нотации и процесса, в котором участвует молекула.

Более детальное описание этой части элемента приведено на рис.6. На этом рисунке прямоугольники обозначают индивиды, а стрелки – свойства. Внутри каждого прямоугольника выписаны свойства и значения свойств, относящихся к данному индивиду. Заметим, что индивид A0 является элементом класса T6-IS, A1 – элементом класса OutputData_MD, A2 – элементом класса TransitionsQuantumNumbers_MD, A3 – элементом класса EinsteinCoefficient_MD, A4 – элементом класса Wavenumbers_MD, B1 B2 – элементами класса BandQuantumNumbersList.

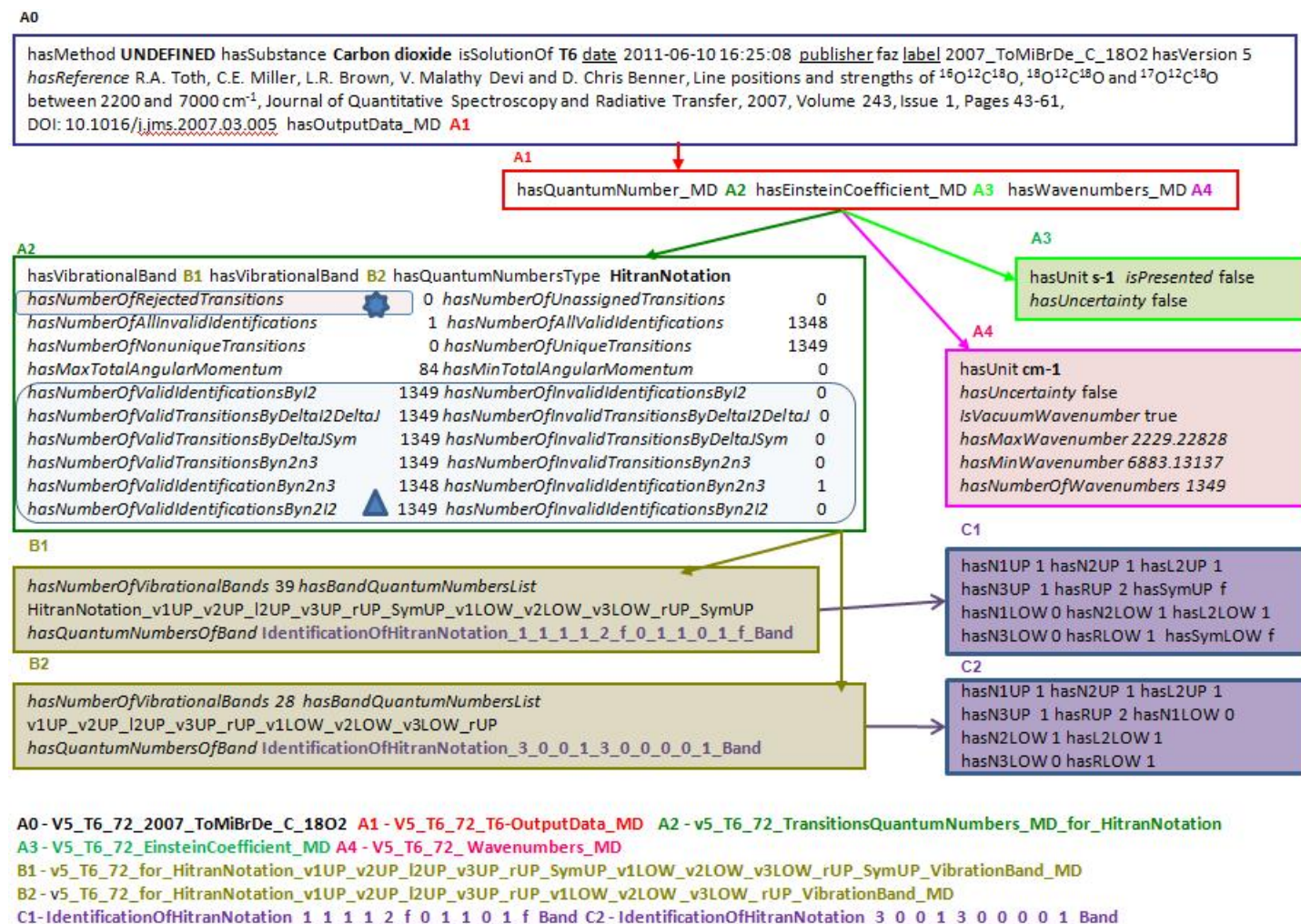


Рис. 6. Структура индивида v5-T6-72-2007_ToMiBrDe_C_18O2 (прикладная онтология версии 5).

4.2. Относительные свойства элемента класса T6-IS

В данной работе рассмотрены относительные свойства только пар элементов одного и того же класса. К числу конкретных свойств пар элементов отнесены свойства, характеризующие стандартные отклонения для двух решений задачи, максимальную разность между идентичными волновыми числами, нарушения порядка следования квантовых чисел при упорядочении численных значений физических величин. Как правило, стандартные отклонения в количественной спектроскопии вычисляются при сравнении расчетных данных и результатов измерений. Более детально процедуры вычислений значений этих свойств рассмотрены в пятой части статьи.

Рассмотрим пример субъектно-предикатной структуры, соответствующей молекуле диоксида углерода, для описания квантовых чисел которой используются нотации Hitran в описаниях спектров поглощения (см. рис.7). Пара источников данных 2007_ToMiBrDe и 1983_EsRo, содержащих вакуумные волновые числа, имеет две одинаково идентифицированные колебательно-вращательные полосы A1 и A2, каждая из которых может описываться двумя разными наборами квантовых чисел. Для каждой соответствующей пары вычислены стандартные отклонения. Полосы содержат часть волновых чисел, поэтому одновременно вычисляется значение свойства, характеризующего стандартное отклонение решений независимо от разбиения на полосы (свойство hasTotalSDValue).

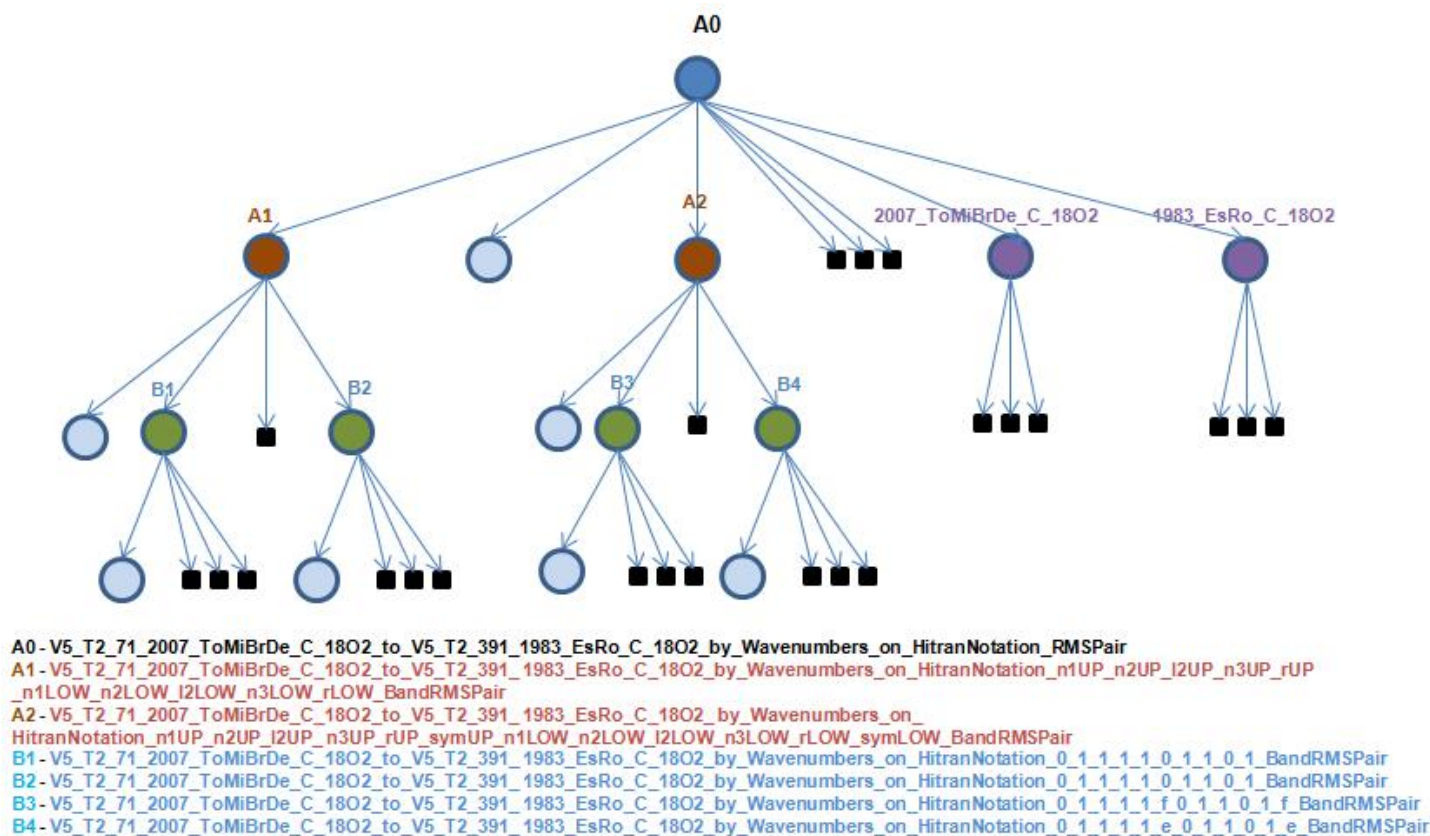


Рис.7. Субъектно-предикатная структура элемента класса RMSPair, включающая часть относительных свойств, связанных со стандартными отклонениями.

Детальное описание этой части элемента приведено на рис.8. Описание этого рисунка использует те же обозначения, что и на рис.6. Индивиды, показанные на рис.7 и 8, являются элементами классов. A0 является элементом класса RMSPair, A1 и A2 являются элементами класса BandRMSPair, B1, B2 - элементы класса RMSVibrationalBand.

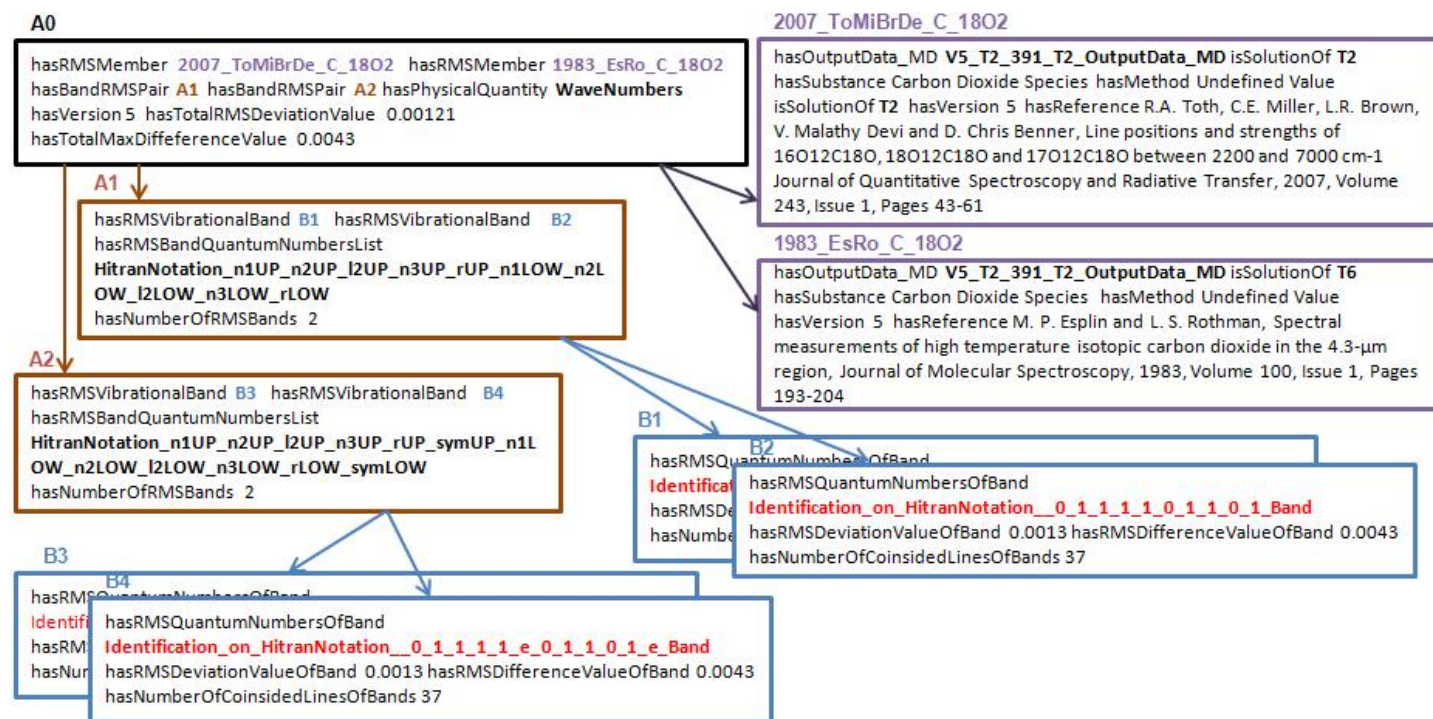


Рис. 8. Структура индивида V5_T2_71_2007_ToMiBrDe_C_18O2_to_V5_T2_391_1983_EsRo_C_18O2_by_Wavenumbers_on_HitranNotation_RMSPair (прикладная онтология версии 5).

Представленные на рис.5-8 структуры характеризуют в деталях представление описания свойств решения задачи T6. Обратимся к рассмотрению аналогий между созданными в нашей работе источниками информации и библиографическими записями функциональные особенности которых описаны в [4,5].

В нашей работе моделью публикации является набор источников данных, содержащихся в статье. Источники информации содержат описание этих источников данных или, другими словами, свойств решений задач спектроскопии. Что общего между источниками информации и библиографическими записями и в чем их различие?

Библиографические записи содержат объекты трех групп. Общим является то, что эти объекты присутствуют и в источниках информации, но степень детализации описания этих объектов различна. В нашей работе для описания внешних атрибутов публикации используется редуцированный набор, как объектов, так и атрибутов объектов первой и второй группы. Более того некоторые объекты опущены, например, объект «организация». Но эти объекты пользователь может восстановить в соответствующих библиотечных системах, используя библиографическую ссылку источника данных. С другой стороны, уровень детализации таких объектов третьей группы как «предмет» и «концепция» в статической части источника информации существенно выше, чем в библиографическом описании, в котором эти объекты практически не атрибутированы [5]. Наконец, основным различием между источником информации и библиографической записью является зависимость значений некоторых свойств, используемых при конструировании источника информации, от времени, а также тот факт, что источник информации формируется в автоматическом режиме. Таким образом, между электронными библиотеками и информационными системами трехслойной архитектуры [30] (приведенная онтология является частью такой системы) существует тесная связь, как в архитектуре, так и в функциональности.

5. Реализация работы с индивидами в машине вывода FaCT++ [a5]

Для формализации знаний используются подходы, основанные на разных формальных логиках. Один из них, широко применяемый в настоящее время, основан на дескриптивной логике (ДЛ) [31]. Формирование подхода началось в 70-х годах, когда были зафиксированы синтаксис и фреймовая семантика. В конце 80-х ДЛ получают теоретико-множественную семантику и (синтактические) алгоритмы рассуждения, основанные на структурах формул. В 90-е годы появляются первые полные системы вывода для достаточно сложных ДЛ [32,33]. В это же время появляются различные *ad-hoc* стандартизации синтаксиса языка. В 2000-х были созданы системы вывода для все более и более выразительных ДЛ, и представление знаний, основанное на ДЛ, стандартизуется консорциумом W3C в языках для описания онтологий OWL/OWL 2 [34]. В настоящее время в Вебе большинство информационных ресурсов, содержащих знания, формализованы с помощью этих языков.

Объектами ДЛ являются классы, свойства (отношения) и индивиды. Моделью является пара непустой домен (множество) и функция интерпретации. Функция интерпретации отображает класс в подмножество домена, свойство – в бинарное отношение между элементами домена, индивид – в элемент домена. Сложные классы и отношения получаются из атомарных путем различных конструкторов (объединение, пересечение, дополнение и др.) Набор конструкторов определяет выразительную силу конкретной ДЛ.

База знаний (БЗ), или онтология – это множество аксиом ДЛ. Аксиома ограничивает множество возможных моделей. Различают аксиомы, описывающие свойства классов, отношений и индивидов. Простейшая аксиома классов утверждает, что некоторый (возможно, сложный) класс включается в другой класс. Другая аксиома устанавливает эквивалентность двух классов. Аксиомы отношений задают транзитивность, симметричность, функциональность и другие свойства различных отношений онтологии. Аксиомы индивидов позволяют указать идентичность или неравенство индивидов, задать отношение между индивидами. Исторически множество всех аксиом классов некоторой БЗ называют TBox, аксиомы индивидов – ABox, а аксиомы отношений – RBox. Набор возможных аксиом также определяет выразительную силу ДЛ.

Система вывода ДЛ позволяет решать некоторые задачи рассуждения в ДЛ. Основные задачи следующие:

1. Непротиворечивость БЗ: существует ли модель, удовлетворяющая всем аксиомам одновременно.
2. Выполнимость класса: существует ли модель БЗ, в которой данный класс интерпретируется непустым множеством.
3. Включение классов: верно ли что в любой модели БЗ интерпретация одного из классов (возможно, сложного) всегда включена (в теоретико-множественном смысле) в интерпретацию другого.
4. Проверка включения индивида: для данного индивида определить, является ли он элементом данного класса. Расширения этой задачи:
 - а. Поиск множества элементов класса: определить все индивиды, которые включены в класс для любой модели БЗ.
 - б. Поиск множества типов индивида: определить все классы, элементом которого он является.
5. Классификация: построить таксономию (граф отношения включения) для всех именованных классов БЗ.

Непротиворечивость БЗ проверяется всегда самым первым действием, потому что с противоречивой БЗ (у которой отсутствует модель) работать нельзя. Наиболее частая причина противоречивости – наличие индивида, который является членом невыполнимого класса. Проверка выполнимости класса обычно производится при необходимости проверить совместность каких-нибудь критериев. Приведем пример, относящийся к прикладной онтологии, описанной выше. Пусть класс **IS-SR** содержит публикации по диоксиду углерода и задачам T2 и T6 в которых все переходы удовлетворяют правилам отбора (если p есть значение свойства *hasNumberOfAllInvalidIdentification*, то это класс источников информации для которых $p=0$). Пусть класс IS-noSR содержит публикации, в которых хотя бы один переход не удовлетворяет правилам отбора. Введем класс **CanonicIS**, определяемый как класс, содержащий публикации, в которых все идентифицированные переходы удовлетворяют правилам отбора и ни один из них не отклонен экспертами (если q есть значение свойства *hasNumberOfRejectedTransitions*, то это класс источников информации для которых $q=0$).

EquivalentClass(**IS-SR**, $p \text{ some int}[\geq 0, < 1]$)

EquivalentClass(**IS-noSR**, $p \text{ some int}[\geq 1]$)

SubClassOf(**CanonicIS**, $q \text{ only int}[\geq 0, < 1]$ and $p \text{ only int}[\geq 0, < 1]$)

В такой онтологии уместно проверить, какие публикации, содержащие переходы, являются каноническими, путем проверки выполнимости конъюнкции **IS-SR** and **CanonicIS**, **IS-noSR** and **CanonicIS**.

Если переписать последнюю аксиому как SubClassOf(**CanonicIS**, $q \text{ some int}[\geq 0, < 1]$ and $p \text{ some int}[\geq 0, < 1]$), то эту же задачу будет решать проверка включений **CanonicIS** [= **IS-SR** и **CanonicIS**]=**IS-noSR**.

Рассмотрим задачу классификации индивидов по категориям. Класс RoVibrationalBand содержит все полосы, данные о которых опубликованы и включены в ИВС. Полосы классифицируют по отношению к нижнему колебательному состоянию. Если нижнее колебательное состояние является основным, то говорят о холодных переходах. В свою очередь холодные переходы содержат два типа

полос: фундаментальные и обертона. Код онтологии, описывающий такую классификацию, приведен ниже:

EquivalentClass(**RoVibrationalBand**, **Hot** or **Fundamental** or **Obertones**)

DisjointClasses(**Hot**, **Fundamental**, **Obertones**)

EquivalentClass(**HotTransition**, **RoVibrationalBand** and *hasQuantumNumbersOfBand* some **Hot**)

EquivalentClass(**ColdTransition**, **RoVibrationalBand** and *hasQuantumNumbersOfBand* some (**Fundamental** or **Obertones**))

Functional(*hasQuantumNumbersOfBand*)

Теперь любой индивид, являющийся элементом класса RoVibrationalBand, может попасть в одну из двух категорий переходов – горячий или холодный. Эта задача дает возможность выяснить, переходы какого типа содержатся в публикациях.

Для достаточно выразительных ДЛ (например, SROIQ) все эти задачи равнозначны [35] и сводятся одна к другой, скажем, к задаче проверки включения индивида (нескольким последовательным задачам в случае, например, задачи классификации). Действительно, БЗ непротиворечива, если некоторый произвольный индивид (не присутствующий в сигнатуре БЗ) является элементом максимального класса (представляющего собой весь домен); некий класс выполним, если добавление аксиомы о включении нового индивида в этот класс не делает БЗ противоречивой; и т.д.

На сегодняшний день существуют несколько систем вывода (для ДЛ такие системы называются ризонерами [reasoner]), позволяющими доказывать утверждения в самом выразительном разрешимом фрагменте стандарта OWL: OWL 2 DL. Среди этих систем можно выделить такие как FaCT++ [36], Hermit [37], RacerPro [38], Pellet [39]. Алгоритмы всех этих систем основываются на той или иной вариации табличного метода. Этот метод зарекомендовал себя хорошей производительностью и расширяемостью в области ДЛ, в большинстве случаев обеспечивая теоретически наилучшие сложности оценки.

Сущность табличного метода [35] состоит в следующем. Известно, что у любого выполнимого класса существует модель. Табличный метод предлагает проверку существования модели путем построения структуры специального вида – таблицы. Таблица – размеченный граф специального вида, удовлетворяющий набору ограничений. Эти ограничения описывают семантику конкретной логики. При этом модель существует в том и только в том случае, если существует таблица, удовлетворяющая соответствующим ограничениям. Таблица представляет собой (возможно, бесконечный) граф, вершины которого помечены классами, а ребра – свойствами. Для каждого индивида из сигнатуры БЗ в таблице существует вершина, помеченная этим индивидом. При проверке класса на выполнимость одна из вершин графа помечается этим классом.

Ограничения на структуру графа зависят от выразительности ДЛ и служат способом фиксации семантики ДЛ. Так, например, если вершина помечена классом, который является конъюнкцией, то каждый конъюнкт тоже должен присутствовать в метке этой вершины. Другой пример: каждым индивидом должна быть помечена только одна вершина. Это ограничение фиксирует семантику индивида. Таблица для логики OWL 2 DL содержит около двух десятков ограничений [35]. Структура таблицы для этой логики достаточно регулярна: таблицу можно разбить на две части. Одна – произвольно связанный граф, соответствующий индивидам БЗ. Другая – деревья (возможно, бесконечные), имеющие корнями вершины этого графа.

Строить таблицу непосредственно в явном виде неудобно и не всегда возможно. Во-первых, таблица может быть бесконечной (в частности, выразительных способностей OWL 2 DL достаточно, чтобы построить выполнимый класс, который имеет только бесконечные модели). Во-вторых, ограничения таблицы имеют декларативный характер, что затрудняет ее конструирование. Поэтому вместо непосредственно таблицы системы вывода ДЛ по алгоритму, основанному на ограничениях таблицы, строят конечную структуру [35], называемую граф вывода (ГВ) (Completion Graph), по которой можно воссоздать модель. Чем сложнее логика (чем больше ее выразительная способность), тем больше разница между таблицей и ГВ.

Опишем реализацию табличного метода, используемую в системе вывода FaCT++, разработанную в работе [36] одним из авторов статьи. Для выразительных ДЛ, с которыми работает FaCT++, табличная структура (ГВ) представляет собой размеченный ориентированный граф специальной формы, вершины которого помечены концептами и/или индивидами, а дуги – свойствами. Вместо ограничений графа используются правила вывода, полученные из них. В частности, для конъюнкции правило имеет следующий вид: если некоторая вершина помечена конъюнкцией классов, но не все конъюнкты принадлежат метке этой вершины, то необходимо добавить недостающие конъюнкты к метке данной вершины. Аналогичным образом преобразуются в правила вывода и другие ограничения таблицы. Форма ГВ в целом отражает форму таблицы. Основу ее составляет произвольный граф, близко соответствующий таковому из таблицы. Вершины этого графа, в свою очередь, являются корнями конечных (и в этом отличие от таблицы) деревьев. Для того, чтобы была возможность представлять бесконечные модели, используется так называемый механизм блокировки.

В силу особенностей OWL 2 DL, единственная возможность построения бесконечных моделей – это многократное повторение фрагмента одного из деревьев, “растущих” из центрального графа. Поэтому, табличный алгоритм, используемый в FaCT++, распознает точки повторения шаблона, и завершает процедуру построения графа в данном листе, помечая его статус как заблокированный. Никакие правила расширения не применяются в заблокированных вершинах.

Еще одна особенность ГВ для языка OWL 2 DL – наличие различных вершин. В силу наличия в этом языке считающих кванторов, возникает необходимость склеивать разные вершины друг с другом и помечать некоторые вершины как различные в момент создания. В FaCT++ склейка вершин производится как копирование всей информации одной из склеиваемых вершин (копия) в другую (мастер), и разметка копии как заблокированной мастером. Для учета различных вершин используется множество списков вершин. В каждом списке находятся вершины, которые нельзя склеивать.

Замечания по реализации. В ГВ, который строит FaCT++, вершины графа, также как и ребра, представляются отдельными программными объектами. Среди их свойств можно выделить метки (множества классов и отношений, соответственно), указатель на блокирующий объект (если таковой имеется) и множество вершин, различных с данной. Перед проверкой выполнимости класса ГВ инициализируется следующим образом. Для каждого индивида строится соответствующая вершина в графе, плюс еще одна (безымянная) вершина для проверяемого класса. Все эти вершины помечаются индивидами; безымянная вершина помечается проверяемым классом.

Помимо этого, осуществляются действия, предпринимаемые по наличию соответствующих аксиом в БЗ:

- SameAs($i_1, \dots, i_n$): все вершины, соответствующие индивидам из аксиомы, склеиваются в одну. В дальнейшем они считаются синонимами и не различаются алгоритмом.
- SameAs($i_1, \dots, i_n$): все вершины, соответствующие индивидам из аксиомы, склеиваются в одну. В дальнейшем они считаются синонимами и не различаются алгоритмом.
- DifferentFrom($i_1, \dots, i_n$): все вершины, соответствующие индивидам, помечаются в графе как различные.
- ClassAssertion(C, i): вершина, соответствующая индивиду из аксиомы, помечается соответствующим классом.
- RelatedIndividuals(R, i, j): между индивидами из аксиомы в графе проводится ребро, помеченное отношением, поименованным в аксиоме.

После такой инициализации выполнение табличного алгоритма происходит стандартным способом: для всех возможных правил проверяется, нет ли вершины, в которой это правило применимо. Если такая вершина находится, правило срабатывает и граф вывода расширяется соответствующим образом.

Интересно, что два индивида, не идентичные в силу аксиомы SameAs, тем не менее могут оказаться идентичными. Этот факт может быть выведен с помощью системы вывода.

Для примера, рассмотрим следующую БЗ.

ObjectPropertyAssertion (R, o, i1);

ObjectPropertyAssertion (R, o, i2);

Functional(R)

Здесь индивид o связан отношением R с индивидами i1 и i2. Но отношение R функционально: оно ведет только к одной точке. Значит, эта точка будет помечена и i1 и i2. То есть, эти индивиды одинаковы.

Заключение

В статье рассмотрена предметная область «Источник информации» и построены базовая и прикладная онтологии источников информации о решениях прямой и обратной задачи определения характеристик переходов изолированной молекулы. Эти онтологии ориентированы на задачу поиска достоверных публикаций о вакуумных волновых числах. В работе достоверность публикации определяется по ряду формальных критериев, к числу которых относятся, в частности, правила отбора. Для молекулы диоксида углерода приведен реестр нотаций и правил отбора. Источник информации характеризуется двумя основными наборами свойств. Индивидуальные свойства характеризуют только публикацию, с которой он связан. Относительные свойства описывают некоторые отношения между источником данных некоторой публикации и всеми источниками данных из других публикаций, относящихся к рассматриваемой молекуле.

Проведено сравнение введенного понятия «источник информации» с понятием «библиографическая запись» [5]. Источник информации обобщает понятие «библиографическая запись» в части включения относительных свойств и свойств зависящих от времени. Источник информации, ориентированный на описание деталей предметной области, существенно детальнее описывает такие объекты библиографической записи как «предмет» и «концепт».

Дано краткое описание задач, которые решаются при работе машины вывода, используемой нами при работе с описанными онтологиями. Сжато описан табличный метод, применяемый в машине вывода FACT++.

Дальнейшее развитие работы направлено на систематизацию реестра правил отбора для атмосферных молекул и реестра нотаций квантовых чисел. Отдельно упомянем о необходимости более детального анализа относительных свойств. В первую очередь, таких как нарушения порядка следования квантовых чисел выявляемого при сравнении упорядоченных по значению вакуумных волновых чисел при сравнении источников данных. Наконец, интерес представляет статистика достоверных и доверительных источников данных (соответственно, публикаций).

Авторы (Привезенцев А., Фазлиев А.) благодарны Российскому фонду фундаментальных исследований за частичное финансирование работы в рамках гранта (11-07-00660-а).

ЛИТЕРАТУРА

1. Лаврентьев Н.А. Базы знаний для описания информационных ресурсов в молекулярной спектроскопии 2. Модель данных в количественной спектроскопии / Лаврентьев Н.А., Привезенцев А.И., Фазлиев А.З. // Электронные библиотеки, 2011, т. 14, в.2. [Электронный ресурс] – <http://elbib.ru/index.phtml?page=elbib/rus/journal/2011/part2>
2. Российский коммуникативный формат представления классификационных данных (российская версия UNIMARC Classification Format). [Электронный ресурс] – <http://www.rusmarc.ru/commun/index.html>
3. MARC 21 Format for Bibliographic Data. [Электронный ресурс] – <http://www.loc.gov/marc/bibliographic>
4. Функциональные требования к библиографическим записям : окончат. отчет / Рос. библ. ассоц., Рос. гос. б-ка ; пер. с англ. [В. В. Арефьев] ; науч. ред. пер.: Т. А. Бахтурина, Н. Н. Каспарова, Н. Ю. Кулыгина. – Москва : РГБ, 2006. – [150] с.
5. Functional Requirements for Bibliographic Records, UBCIM Publications – New Series Vol 19, Final Report, 1998. [Электронный ресурс] – www.ifla.org/files/cataloguing/frbr/frbr.pdf
6. Лукашевич Н.В., Тезаурусы в задачах информационного поиска, М.: Из-во МГУ, 2011, 512С.
7. Oberle, D. Semantic management of middleware, Berlin: Springer, 2006. 268 pp.
8. Privezentsev A. Computed Knowledge Base for Description of Information Resources of Water Spectroscopy / Privezentsev A., Fazliev A., Tsarkov D., Tennyson J. // Proc. of the 7th International Workshop on OWL: Experiences and Directions (OWLED 2010), San Francisco, California, USA, June 21-22, 2010. Edited by Evren Sirin, Kendall Clark, CEUR-WS Proc. Vol-614, [Электронный ресурс] – <http://ceur-ws.org/Vol-614/>

9. Lavrentiev N.A. Complete set of published spectral data on CO₂ Molecule / Lavrentiev N.A., Privesentsev A.I., Fazliev A.Z., Filippov N.N. // Abstracts of the 22-nd Colloquium on High Resolution Molecular Spectroscopy, 2011, p.353.
10. Voronina S.S. Systematization of the published spectroscopic parameters of ammonia / Voronina S.S., Yurchenko S.N., Fazliev A.Z. // Abstracts of the 22-nd Colloquium on High Resolution Molecular Spectroscopy, 2011, p.163
11. Козодоев А.В. Систематизация опубликованных параметров спектральных линий молекул метана, силана и германа / Козодоев А.В., Вельмузова И.А., Сенников П.Г., Фазлиев А.З., Филиппов Н.Н., Григорович Н.М. // Сборник тезисов Международного симпозиума «Атмосферная радиация и динамика» (МСАРД - 2011) Санкт - Петербург, с.102-103.
12. Половцева Е.Р. Информационная система для решения задач молекулярной спектроскопии. 5. Колебательно-вращательные переходы и уровни энергии молекулы H₂S / Половцева Е.Р., Лаврентьев Н.А., Воронина С.С., Науменко О.В., Фазлиев А.З. // Оптика атм. и океана. 2011, Т.24, №10, с. 898-905
13. Козодоев А.В. Аннотирование информационных ресурсов в распределенной информационной системе "Молекулярная спектроскопия" / Козодоев А.В., Привезенцев А.И. Фазлиев А.З. // Электронные библиотеки, 2006, т. 9, в.3. [Электронный ресурс]: - <http://www.elbib.ru/index.phtml?page=elbib/rus/journal/2006/part3/KPF>
14. Зиновьев А.А. Основы логической теории знаний. М.: Наука, 1967. 260 с.
15. OWL Web Ontology Language Semantics and Abstract Syntax W3C Recommendation 10 February 2004. [Электронный ресурс] - <http://www.w3.org/TR/2004/REC-owl-semantics-20040210/>
16. Naus H. Cavity -Ring-Down Spectroscopy on Water Vapor in the Range 555-604 nm / Naus H., Ubachs W., Levelt P. F., Polyansky O. L., Zobov N. F. and Tennyson J. // J. Mol. Spectrosc., 2001, v.205, p.117-121.
17. Sollins K. Functional Requirements for Uniform Resource Names / Sollins K., Masinter L. // RFC 1737, 1994. [Электронный ресурс] - <http://tools.ietf.org/html/rfc1737>
18. Berners-Lee T. URI Generic Syntax / Berners-Lee T., Fielding R., Masinter L. // RFC 3986, 2005. [Электронный ресурс] - <http://tools.ietf.org/html/rfc3986>
19. Dubernet M.L. Virtual atomic and molecular data centre / Dubernet M.L., Boudon V., Culhane L. et al. // J. Quant. Spectrosc. & Rad. Transfer. 2010. v. 111, No 15. p. 2151-2159.
20. Jacquinet-Husson N. The GEISA spectroscopic database: current and future archive for earth and planetary atmosphere studies / Jacquinet-Husson N., Scott N.A., Chedin A. et al. // J. Quant. Spectrosc. & Rad. Transfer. 2008. v. 109. No 6. p. 1043-1059.
21. Tashkun S.A. CDSD-4000: High-Temperature Spectroscopic CO₂ Databank / Tashkun S.A. and Perevalov V.I. // The 11th HITRAN Database Conference, June 16 - June 18, 2010, Cambridge, 2010, p. 10.
22. Toth R.A., Spectroscopic database of CO₂ line parameters: 4300-7000 cm⁻¹ / Toth R.A., Brown L.R., Miller C.E., Devi V. Malathy and Benner D.Chris // J. Quant. Spectrosc. & Rad. Transfer.2008, v. 109, No 6, p. 906-921.
23. Rothman L.S. The HITRAN 2008 molecular spectroscopic database / Rothman L.S., Gordon I.E., Barbe A. et al. // J. Quant. Spectrosc. & Rad.Transfer. 2009. v. 110, No 9. p. 533-572.
24. Sowa J.F. Knowledge representation: logical, philosophical and computational foundations / Sowa J.F. - Pacific Grove, CA: Brooks Cole, 2000. 594 p.
25. Van Welie, M. An Ontology for Task World Models / M. van Welie, G. C. van der Veer, A. Eliens // 5th International Eurographics Workshop on Design Specification and Verification of Interactive Systems - Abingdon, 3-5 june 1998. - 14pp.
26. Chandrasekaran, B. Ontology of tasks and methods / B. Chandrasekaran, J. R. Josephson, V. R. Benjamins // Banff Knowledge Acquisition Workshop. -1998. [Электронный ресурс]: <http://www.cse.ohio-state.edu/~chandra/Ontology-of-Tasks-Methods.PDF.-25.pp>.
27. OWL Web Ontology Language Reference, W3C Recommendation 10 February 2004 [Электронный ресурс]: <http://www.w3.org/TR/2004/REC-owl-ref-20040210/>
28. Лаврентьев Н.А. Сравнение спектральных массивов данных HITRAN и GEISA с учетом ограничения на опубликование спектральных данных / Лаврентьев Н.А., Макогон М.М., А.З.Фазлиев // Оптика атм. и океана. 2011. т.24. №4. с.279-292.
29. ИВС W@DIS, [Электронный ресурс] - <http://wadis.iao.ru>.
30. Ахлестин А.Ю. Информационная система трехслойной архитектуры / Ахлестин А.Ю., Лаврентьев Н.А., Привезенцев А.И. Фазлиев А.З. // Труды семинара «Наукоемкое программное обеспечение». 2011. Новосибирск, с.38-43.
31. The Description Logic Handbook: Theory, Implementation, and Applications / Baader F., Calvanese D., McGuinness D., Nardi D., Patel-Schneider P. (eds.) // Cambridge University Press, second edn. (2007).
32. Haarslev, V. Consistency testing: The RACE experience / Haarslev, V., Moller, R. // In: Proceedings, Automated Reasoning with Analytic. pp. 57-61. Springer-Verlag, 2000.
33. Horrocks I. DL systems comparison / Horrocks I., Patel-Schneider, P.F. // In: Proc. 1998 Int. Workshop on Description Logics (DL'98). vol. 11, pp. 55-57, 1998.
34. OWL Working Group, W.: OWL 2 Web Ontology Language: Document Overview. W3C Recommendation (27 October 2009), [Электронный

ресурс] - <http://www.w3.org/TR/owl2-overview/>

35. Horrocks I. The even more irresistible SROIQ / Horrocks I., Kutz O., Sattler U. // In: Proc. 10th Int. Conf. on Principles of Knowledge Representation and Reasoning (KR'06). pp. 57-67. AAAI Press, 2006.

36. Tsarkov D. FaCT++ description logic reasoner: System description / Tsarkov, D., Horrocks I. // In: Proc. 3rd Int. Joint Conf. on Automated Reasoning (IJCAR'06). LNCS, vol. 4130, pp. 292-297. Springer, 2006.

37. HerMiT Reasoner, [Электронный ресурс] - <http://www.hermit-reasoner.com>

38. Haarslev V. The RacerPro knowledge representation and reasoning system / Haarslev V., Hidde K., Moller R., Wessel M. // Semantic Web, 2012, accepted for publication.

39. Sirin E., Pellet: A practical OWL-DL reasoner / Sirin E., Parsia B., Grau B.C., Kalyanpur A., Katz Y. // J. of Web Semantics 5(2), 51-53, 2007.

Об авторах

Привезенцев Алексей Иванович - к.т.н., научный сотрудник Центра интегрированных информационных систем Института оптики атмосферы им. В.Зуева СО РАН, E-mail: remake@iao.ru;

Царьков Дмитрий Викторович - к.ф.-м.н., научный сотрудник, университет Манчестера, Великобритания, E-mail: tsarkov@cs.man.ac.uk

Фазлиев Александр Зарипович - к.ф.-м.н., заведующий Центром интегрированных информационных систем Института оптики атмосферы им. В.Зуева СО РАН, E-mail: faz@iao.ru