# Цифровая библиотека научных статей по количественной спектроскопии

© З.В. Апанович © П.С. Винокуров
Институт систем информатики СО РАН
Новосибирск
apanovich@iis.nsk.su vinokurov.pasha@gmail.ru

© А.Ю.Ахлёстин © А.И. Привезенцев © А.З.Фазлиев
Институт оптики атмосферы СО РАН
Томск
lexa@iao.ru faz@iao.ru

**Аннотация**

В докладе обсуждается подход к построению цифровой библиотеки научных статей для предметной области, в которой фактологическая часть существенно превосходит понятийную. На примере библиотеки статей по количественной спектроскопии показано как использование модели публикации (статьи), содержащей решение задач предметной области и свойства этого решения, приводит к автоматической каталогизации решений.

Особое внимание в работе уделено визуализации индивидов онтологии, характеризующих пары источников информации. Визуализация позволяет давать качественную оценку состоятельности решений задач спектроскопии в случае анализа данных большого объема.

*Авторы благодарны РФФИ (гранты 11-07-00660 и 11-07-0038) и РАН (проект РАН 15/10) за финансирование работы.*

## 1 Введение

Систематизация ресурсов библиотечного фонда в библиотеках основана на библиографических записях, определяемых тем или иным стандартом [1]. В большинстве библиотек научные статьи не являются единицами хранения. В цифровых библиотеках научные публикации в журналах, трудах конференций или сборниках статей являются единицами хранения, и, как правило, в таких библиотеках содержатся системы поиска статей, основанные на библиографических записях, относящихся к статьям. В работе [2, 3] описаны функциональные требования к библиотечным записям.

**Труды 14-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» — RCDL-2012, Переславль-Залесский, Россия, 15-18 октября 2012 г.**

На практике системы поиска ресурсов в научных цифровых библиотеках опираются на тексты статей на естественном языке. Большая часть поисковых систем не использует формализованные понятия предметных областей, содержащиеся в искомых ресурса в явном или неявном виде, по простой причине: большинство понятий не формализовано. Задачу поиска с учетом терминологии предметной области решают с помощью информационно-поисковых тезаурусов (см., например, [4]). 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 [5] характеризуют наиболее динамичную часть знаний и требуют для их представления описания с большим количеством деталей.

В докладе рассматривается онтология источников информации, связанных с публикациями по 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 и параметров спектральных линий.

При выполнении грантов и проектов авторы собрали опубликованные решения шести задач в количественной спектроскопии для атмосферных молекул (вода [6], диоксид углерода [7], аммиак [8], метан [9], сероводород [10] 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 в работе [11].

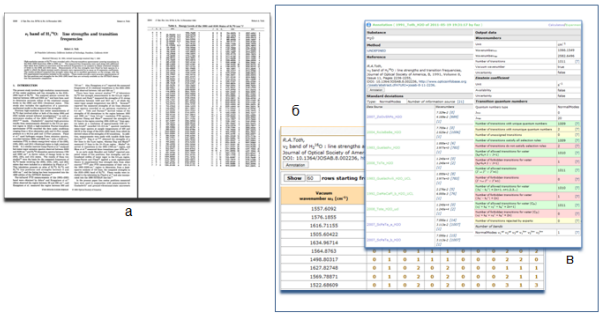


Рис.1 *Модель публикации. а) – оригинал публикации, б) – решение задачи спектроскопии и в) – свойства решения задачи спектроскопии.*

Каждое такое решение является частью публикации, а в количественной спектроскопии основной частью, так как содержит наибольшее число типизованных фактов. В простом случае выбирая модель публикации или ее части, можно ограничиться решением задач спектроскопии, что соответствует публикации таких решений на сайтах и FTP. 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 [12].

Ниже рассмотрена модель публикации для предметной области «Количественная спектроскопия». Она может применяться и для других предметных областей в которых фактологическая часть значительно превышает понятийную. Эта модель позволяет автоматизировать процесс каталогизации научных статей и их частей. Рассмотрен пример библиотеки по количественной спектроскопии. Наконец, представлены примеры визуализации индивидов, характеризующих индивидуальные свойства решений задач спектроскопии и парных отношений между источниками данных.

Целью выполненной работы являлось построение цифровой библиотеки, в рамках которой возможен дифференцированный поиск достоверных информационных ресурсов или недостоверных ресурсов в предметной области по принятым в ней критериям. Для достижения цели построена модель публикации, выбраны языки спецификации данных, информации и знаний, созданы источники данных и информации. Показано, что анализ достоверности значительного количества источников данных существенно проще при графическом представлении парных отношений.

## 2 Модель публикации

Создание модели публикации для цифровой библиотеки научных статей связано с задачей автоматической каталогизации информационных ресурсов по количественной спектроскопии. Имеющаяся у авторов коллекция статей уже превышает 8000 публикаций, относящихся к периоду с 1926 года по настоящее время. Выбранная модель предметной области [13] содержит решения шести задач спектроскопии, имеющих определяющее значение для прикладных предметных областей таких как астрономия, оптика атмосферы, спектроскопия и т.д.

Основой для построения модели является допущение о том, что публикация содержит факты, являющиеся решениями ряда задач спектроскопии. Эти факты разделены на две группы. К первой группе относятся такие решения задач, которые можно отнести к одной молекуле и одному методу решения. Ко второй группе относятся все оставшиеся решения.

Извлеченные из публикаций факты являются частями и представляются в цифровой библиотеке в форме первичных или составных источников данных. С каждым источником данных связывается источник информации, содержащий свойства соответствующего решения задачи спектроскопии. На рис.1 схематически показано представление модели в виде двух частей с помощью интерфейсов представления данных и информации.

Для формирования источника данных используется реляционная модель данных, а источника информации – язык онтологий OWL DL. Представление источников информации в виде индивидов онтологии является основой для автоматической каталогизации решений задач спектроскопии.

Рис.1 б) демонстрирует отображение данных соответствующей публикации в информационной системе, а рис.1 в) – представление автоматически сгенерированных в ИС свойств этого решения.

### 2.1 Независимые части публикации (первичные источники данных)

Разнообразие молекул, для которых решались задачи, выделенные в работе [11], 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 или неверные решения. 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 [14] (*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*. Значение этого свойства должно быть определено явно и являться публикацией.

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 [15,16], России [17], США [18,19] и т.д.

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

### 2.2 Составные источники данных (агрегации первичных данных в статьях)

*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 [19]).

### 2.3 Источник информации

Первичный источник можно наделять дополнительными свойствами. Перечень и число этих свойств зависит от информационных задач, для решения которых 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

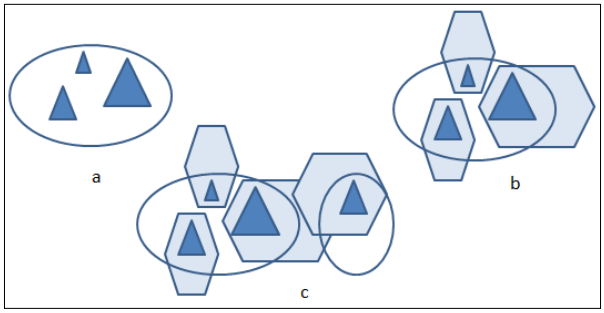


Рис.2 *Соотношения между публикациями, источниками данных и источниками информации.*

свойства, значения которых вычисляются автоматически. Как правило, источник информации включает в себя некоторые высказывания из публикации, содержащей источник данных, который этот источник информации описывает. Большая часть источника информации характеризует знания, содержащиеся 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.

В данной работе изучаются конкретные источники информации, всегда связанные с источником данных (другими словами кардинальность свойства *hasDataSource*, доменом которого является класс источников информации, равна 1).

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

## 3 Цифровая библиотека научных статей

Накопление и распространение опубликованных научных статей в значительной мере ограничено законодательством разных стран. Это обстоятельство приводит к тому, что для значительной части исследователей научные факты недоступны. С появлением сети Интернет проблема доступа к

опубликованным данным постепенно решается. Это связано с тем, что для большинства исследователей интерес представляют фактологические части публикаций, например, результаты измерений в естественных науках. Подобные наборы фактов все чаще накапливаются в базах данных, доступных в сети Интернет.

Однако, в большинстве случаев, целью сбора данных является создание экспертных массивов. Такой подход характерен для исследователей, работающих в прикладных, по отношению к фундаментальным научным дисциплинам, областях.

Библиотечная деятельность ориентирована на организацию всей содержащейся в статьях информации в формах, удобных для исследователя. Как правило, эта деятельность ограничена поддержкой систем поиска информации.

На наш взгляд цифровые научные библиотеки могут сосредоточиться на автоматической интеграции цифровой информации, связанной с научными публикациями. Такая интеграция должна быть тесно связанной с решением задачи проверки непротиворечивости интегрируемой информации.

Решение задачи интеграции требует переосмысления используемых на практике структур библиографической записи, а, следовательно, и уточнения функциональных требований к ним.

### 3.1 Библиографическое описание – источник информации

В начале 90-х в Швеции состоялся семинар по библиографическим записям. Одним из результатов семинара была резолюция об определении функциональных требований к библиографическим записям. Созданная позже модель такой записи была попыткой формирования логической основы понимания правил библиографического описания.

В документе функциональные требования к записям [2] определены с точки зрения следующих основополагающих задач пользователей при поиске и использовании библиотечных каталогов:

- «использование данных для того, чтобы **найти** материалы, которые соответствуют заявленным поисковым критериям (например, в контексте поиска всех документов на данную тему или пластинки, выпущенной под конкретным заглавием);
- использование полученных данных для того, чтобы **идентифицировать** объект (например, для подтверждения соответствия документа, зарегистрированного в записи, документу, который искал пользователь, или для обнаружения различий между двумя текстами или пластинками с одинаковым заглавием);
- использование данных, чтобы **выбрать** объект, который отвечает потребностям пользователя (например, выбрать текст на языке, который пользователь знает или вариант компьютерной программы, совместимой с компьютером и операционной системой, доступной пользователю);
- использование данных для того, чтобы приобрести или **получить** доступ к описанному объекту (например, для размещения заказа на покупку издания, передачи запроса копию книги из библиотечной коллекции или чтобы получить онлайновый доступ к электронному документу, хранящемуся в удаленном компьютере)».

Подобные требования можно распространить на источники информации, введенные выше. Они в предложенной модели публикации играют роль библиографической записи, содержа в себе значительную часть свойств, присущих ей. Однако источники информации также содержат то, что не присуще библиографическим записям.

В первую очередь речь идет о свойствах ориентированных на описание качества данных, размещенных в статьях и о корреляциях данных, извлеченных из разных публикаций.

### 3.2 Каталоги – таксономии классов онтологии информационных ресурсов предметной области

Исследование функционального назначения библиографической записи предназначалось для облегчения работы при автоматизации процесса каталогизации информационных ресурсов.

С другой стороны, возникший почти на десятилетие позже подход Semantic Web ориентирован на более широкий круг задач систематизации ресурсов в глобальной информационной системе (Web). Более того, при реализации подхода были созданы соответствующие средства для представления ресурсов с разной степенью детализации. Эта детализация позволяет строить в автоматическом режиме таксономии классов, а машина вывода позволяет отслеживать наследственность и противоречия, возникающие при создании таких таксономий.

Выделенные концепты предметной области в фактологической части статей, включенные в модель публикации, позволяют строить их онтологическое описание. Эти описания наряду с индивидами содержат понятийную часть, представляемую таксономиями классов.

В рамках языка онтологий OWL DL можно строить классы, накладывая ограничения на свойства. Несложно построить в автоматическом режиме все классы по ограничениям на объектные свойства, т.к. число индивидов ограничено.

На рис. 3 показан пример визуализации части таксономии классов, характеризующий типы задач спектроскопии, решения которых описываются прикладной онтологией, и индивидов, связанных с этими классами.

## 4 Пример цифровой библиотеки. ИВС W@DIS

Примером цифровой библиотеки, использующей модель публикации, описанную выше, является информационная система W@DIS. Эта система основана на коллекции публикаций по количественной спектроскопии. В настоящее время в коллекцию входит около восьми тысяч публикаций. Большая часть этих публикаций не может быть выложена в свободный доступ.

Для решения ряда задач предметных областей, связанных со спектроскопией, пользователям необходима только часть фактов, содержащихся в публикациях этой коллекции. Эти факты относятся к решениям шести задач спектроскопии, связанных с нахождением параметров состояний и переходов молекул. Отметим, что коллекция предназначена для исследователей, занимающихся атмосферной спектроскопией.

Оцифрованные факты из публикаций импортированы в информационную систему и представляют собой разные типы источников данных. При импорте данных [20] в систему для каждого источника данных автоматически создается источник информации [14], содержащий описание свойств импортированных решений задач спектроскопии.

Поскольку каждая конкретная публикация представляется в виде набора источников данных и частей источников информации, относящихся к каждому из источников данных, то каталогизацию публикаций можно заменить более детализированной каталогизацией источников информации.

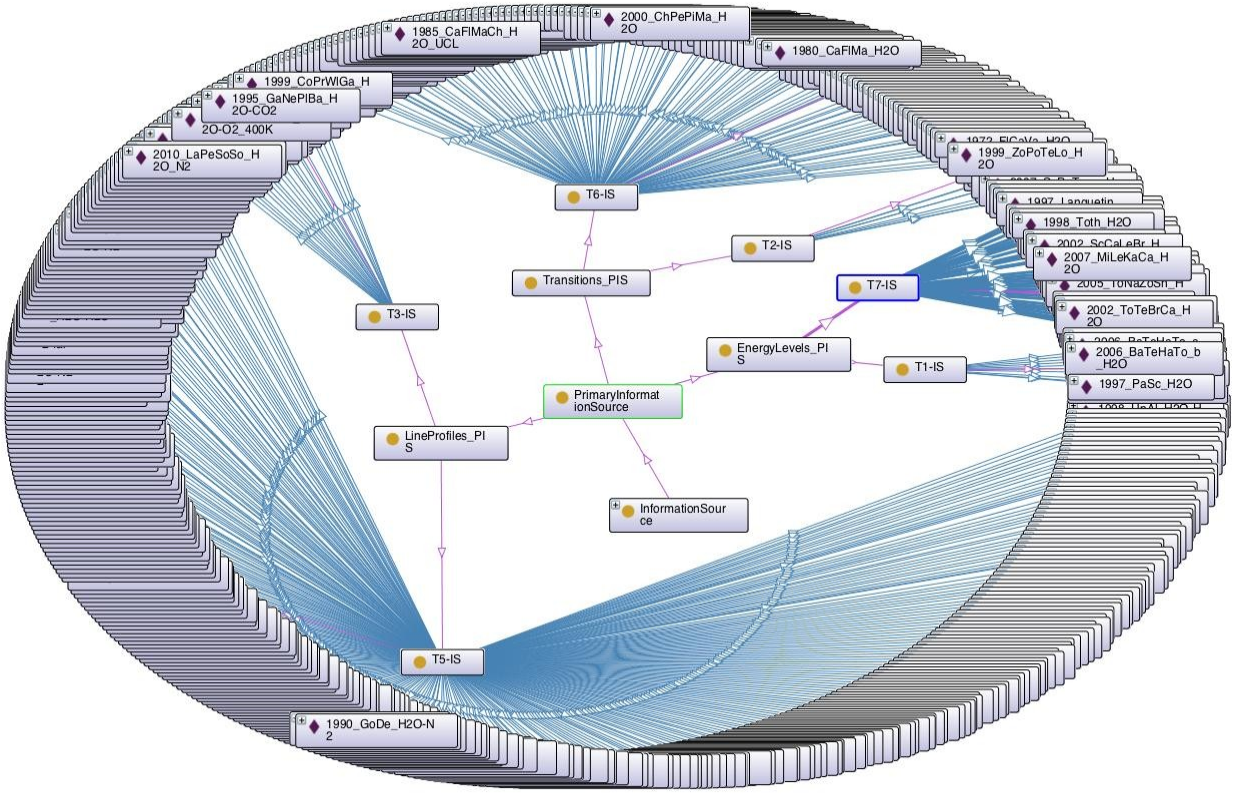


Рис.3 *Представление таксономии классов и относящихся к ним источников информации*

### 4.1 Визуализация индивидов онтологии

Источники информации в ИС представлены с помощью языка онтологий OWL DL. Их можно разделить на две группы. К группе независимых источников отнести те индивиды, которые характеризуют свойства отдельного источника данных, а к другой группе, отнести индивиды, описывающие свойства пары источников данных.

Для работы исследователя с этими индивидами необходимы инструментальные средства визуализации индивидов обеих групп. Ниже подобная визуализация обсуждается на примерах.

Визуализация отдельного источника информации (здесь рассмотрена только визуализация свойств решений задач) позволяет оценить уровень детализации и информационные аспекты анализа данных, характеризуемых этими свойствами. Прежде всего это относится к грубым ошибкам, относящимся к несоблюдению правил отбора.

Визуализация индивида, описывающего свойства пары источников информации, предоставляет более детальную информацию о качестве анализируемых

данных. Она позволяет определять рассогласование между данными, полученными разными группами исследователей по ряду критериев (максимально допустимая разница значений физических величин, среднеквадратическое отклонение, нарушение порядка следования сравниваемых значений). В количественной спектроскопии необходимость визуализации отношений между источниками информации обусловлена огромным количеством значений свойств, используемых при анализе сравниваемых данных.

### 4.2. Независимые источники информации

Демонстрируемые ниже примеры требуют детального описания онтологии, которое приведено в работе [14]. Опуская детали, сосредоточимся на представлении структуры индивида, характеризующего индивидуальные свойства описываемого источника данных и представляющего источник информации. Заметим, что число узлов и листьев остается неизменным при представлении такого индивида в виде дерева.

Проиллюстрируем это на примере. На рис. 4 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.

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.

Индивид, представленный на рис.4, является библиографической записью, относящейся к решению задачи о параметрах спектральных линий, извлеченному из статьи [21].

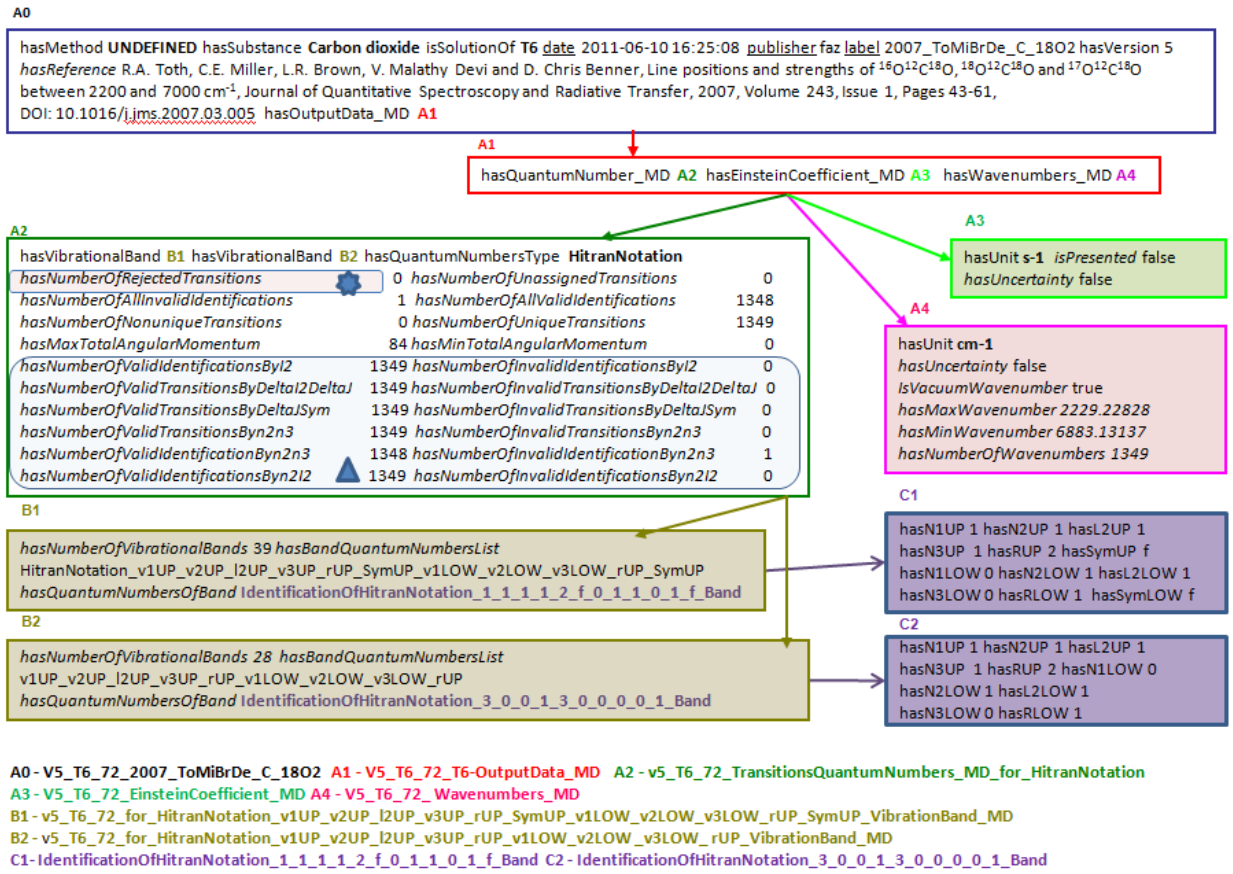


Рис.4 *Представление индивида, характеризующего свойства решения обратной задачи по определению параметров контура спектральной линии*

### 4.3 Источники информации, относящиеся к парам источников данных

Представление источника информации, характеризующего свойства всех пар, включающих выбранный источник данных со всеми другими источниками данных, значительно сложнее. Визуализация такого источника информации для исследователя необходима по ряду причин. Во-первых, в спектроскопии принято сравнивать результаты экспериментов, выполненных разными группами. Во-вторых, таких парных отношений может быть нескольких типов. В-третьих, число источников данных в ИС изменяется во времени (появляются новые работы

по измерению параметров состояний и переходов). В четвертых, увеличивается точность измерений, следовательно, необходим пересмотр количественных значений критериев, определяющих достоверность фактов. В-пятых, число фактов при сравнении источников данных может составлять десятки тысяч, что заставляет представлять их в графическом виде. Пр едставление этой информации в текстовом виде громоздко и позволяет увидеть только локальную картину.

С другой стороны, методы визуализации информации являются общепризнанным инструментом представления глобального взгляда на абстрактные данные большого объема.

Для того чтобы представить реальную картину, дающую представление обо всех публикациях, принадлежащих данному набору и о связях между этими публикациями, необходимо построить и визуализировать графовую модель имеющихся данных.

С этой целью генерировались графы, вершинами которых являются отдельные публикации, а ребрами свойства парных отношений, а затем осуществлялась визуализация этих графов. Прежде всего, следует отметить, что хотя, количество вершин в этих графах не очень велико, эти графы имеют весьма высокую плотность. Плотность графа определяется как отношение количества ребер в данном графе, к количеству ребер полного графа с тем же множеством вершин. Известно, что основной проблемой при визуализации таких графов является большое количество пересечений ребер, которое слабо зависит от выбранного алгоритма визуализации. То есть независимо от алгоритма визуализации, количество пересечений ребер, а значит, и визуальная загруженность изображения велики. Известны подходы, когда для визуализации таких графов используются алгоритмы типа LinLog [22], при котором сильно связанные вершины располагаются близко друг к другу, а слабо связанные вершины – далеко. Ребра графа при этом подходе вообще не изображаются, чтобы не загромождать изображение. В нашем случае такое решение абсолютно непригодно, потому что основная информация, которая интересует пользователя – это именно ребра. Пользователь заинтересован увидеть с первого взгляда, какое ребро (то есть, характеристика парного отношения) представляет собой достоверную информацию, то есть значения, соответствующие этой паре для колебательно-вращательных полос, а также увидеть, насколько эти значения совпадают (или не совпадают). По этой же причине для визуализации данных этого типа не подходит и метод создания жгутов ребер [23]. Поэтому основное внимание при выборе способа визуализации уделялось именно информации, связанной с количеством «хороших» или «плохих» колебательно-вращательных полос.

Как уже было сказано, в качестве вершин графа рассматриваются отдельные публикации. На изображении каждая публикация идентифицируется номером в базе данных, годом публикации и первыми буквами фамилий авторов. Цвет вершины соответствует типу задачи, решение которой описано в данной публикации, а радиус вершины – ее степени, то есть количеству ребер, связывающих эту публикацию с другими публикациями. Поскольку разброс в степенях вершин может быть весьма велик, для вычисления радиуса вершины используется логарифмическая зависимость от ее степени.

Помимо элемента ds:RMSPair, пару публикаций может описывать один элемент ds:BandRMSPair, имеющий те же самые идентификаторы для публикаций и сообщающий, сколько общих «колебательно-вращательных полос» == ds:RMSVibrationalBand имеется в указанных двух публикациях.

Каждая общая колебательная полоса описывается элементом ds:RMSVibrationalBand, в котором кроме авторов информации о парных публикациях есть идентификатор полосы, состоящий из шести чисел (например _1_0_0_0_0_0_ ), квантовые числа.

В описании каждой общей полосы
ds:RMSVibrationalBand
важны два свойства:
ds:hasMaxDifferenceValueOfBand
(максимальная разность значений) и
ds:hasRMSDeviationValueOfBand
(значение отклонения полосы)

Если hasMaxDifferenceValueOfBand > 0.05, или ds:hasRMSDeviationValueOfBand > 0.1, это может указывать на ошибку в данных.

При этом, оценка максимальной разницы значений hasMaxDifferenceValueOfBand является более грубой, оценка максимального отклонения - более тонкой. Поэтому строится как минимум два разных изображения графа, одно изображение соответствует «плохим» максимальным разностям, а второе – «плохим» отклонениям.

Так же как в случае с радиусами вершин, разброс между количеством колебательно-вращательных полос, общих для двух публикаций, тоже весьма велик и может меняться в диапазоне от одной полосы до миллиона. Понятно, что ширина ребра должна каким-то образом зависеть от этого количества полос, но пропорциональная зависимость при таком большом разбросе не совсем уместна, изображение и так загромождено большим количеством ребер. Поэтому все множество ребер сортируется по количеству колебательно-вращательных полос и вводится шкала, состоящая из пяти градаций, и каждое ребро попадает в один из классов. Ребру при визуализации приписывается ширина, соответствующая его классу. Таким образом, ширина ребра соответствует количеству общих колебательно-вращательных полос для одной RMSPair.

Помимо общего количества колебательно-вращательных полос, изображение должно показывать, сколько полос от общего количества

имеют либо «плохую» разность, либо «плохое» отклонение. Для демонстрации этого свойства используется градиентная раскраска каждого ребра.

Если «плохих» ребер нет вообще, то ребро разбивается на три части. В центре прозрачная часть, по краям – серая. Прозрачная часть используется для уменьшения визуальной загруженности изображения.

Если есть «плохие» ребра, каждое такое ребро разбивается на 5 частей, центральная часть – прозрачная, две крайние, самые ближние к инцидентным вершинам части, соответствуют «плохим» полосам. Ближе к центру расположены две симметричные серые части, соответствующие «хорошим» полосам. Длина красной части пропорциональна количеству «плохих» полос.

Что касается собственно алгоритма визуализации, нами была сделана модификация алгоритма Фрюхтермана-Рейнгольда. Дело в том, что силовые алгоритмы такие как алгоритм Фрюхтерман-Рейнгольда [24] или Камада–Кавая [25] не применим к размещению вершин высокой степени в центре изображения. Для графов, соответствующих данной проблеме, такое решение непригодно, поскольку центральная часть и так загружена большим количеством ребер. Поэтому наша модификация алгоритма «выталкивает» вершины высокой степени на периферию изображения, позволяя проследить ребра, инцидентные этим вершинам.

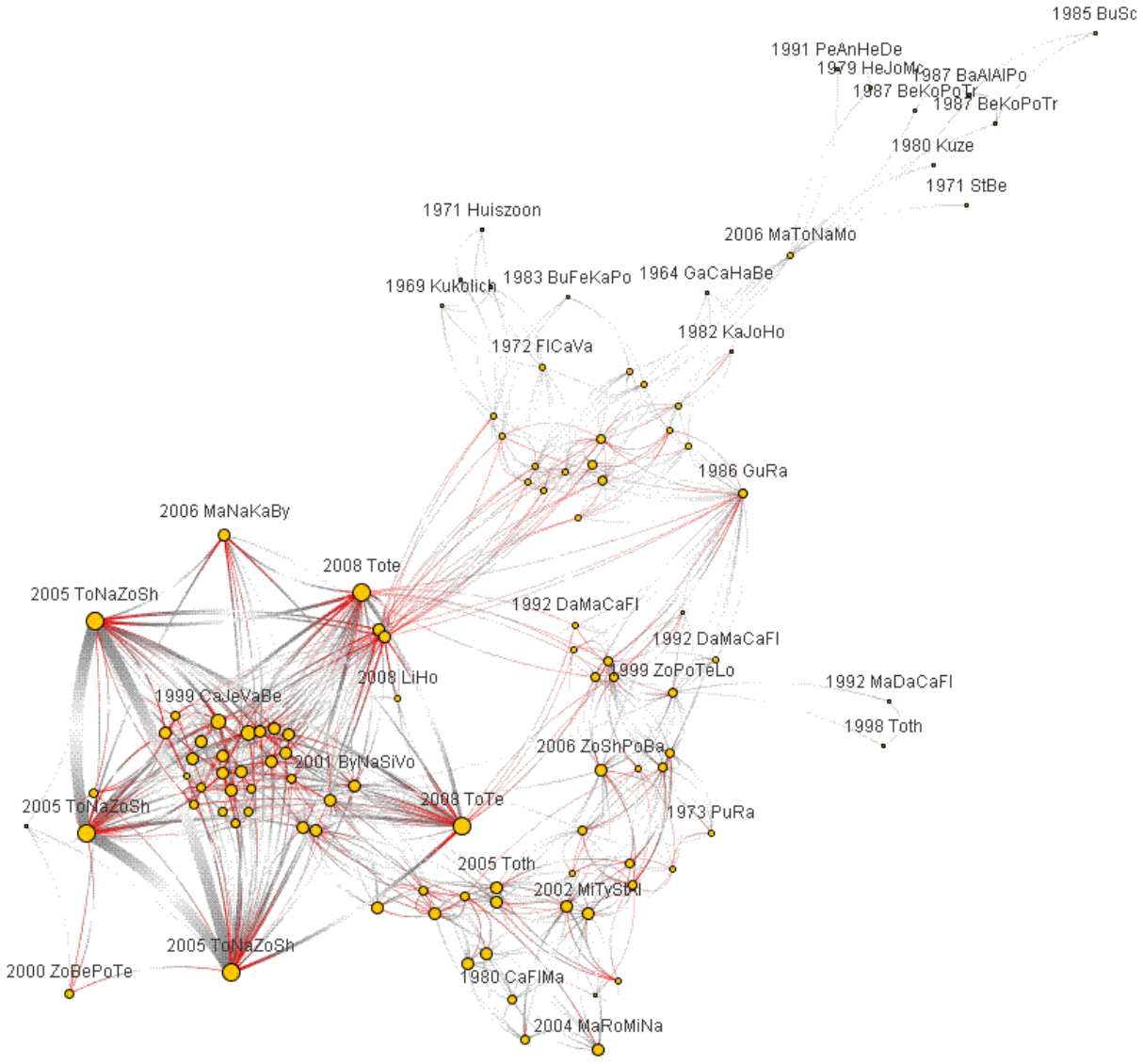


Рис.5 *Графическое представление парных отношений (среднеквадратическое отклонение между отдельными колебательными полосами) между источниками информации на примере первичных источников данных, содержащих решение обратной задачи Т6 [26] для молекулы воды*

Остановимся на интерпретации графа изображенного на рис.5. Узлами графа являются источники информации. Цвет ребер соответствует одному из двух вариантов: красный цвет соответствует неудовлетворительному значению отклонения, а серый цвет – удовлетворительному значению. У сравниваемых источников информации могут содержаться несколько отдельных идентичных колеба-

тельных полос. В таком случае ширина линии, изображающей ребро, становится по величине пропорциональной числу идентичных колебательных полос, а ребро раскрашивается в соответствующие цвета пропорционально числу удовлетворительных или неудовлетворительных значений.

Как следует из цветовой гаммы графа, представленного на рис.5 в публикациях включенных в ИВС существует значительное число переходов молекулы воды среднеквадратические значения которых являются неудовлетворительными.

## 4 Заключение

В работе рассмотрен пример построения цифровой научной библиотеки публикаций. Основное внимание было уделено описанию модели публикации в такой библиотеке. Предложено создавать модели публикаций, содержащих количественную информацию, состоящими из двух частей: результатов решений задач и свойств этих решений. Представление свойств решений в форме индивидов прикладной онтологии позволяет автоматически строить детализированные таксономии классов, главным образом по ограничениям на свойства онтологии. Существенным является то, что при построении таксономий пустые классы в них не включаются.

Согласование фактологических частей публикаций осуществляется по выбранному набору свойств решений задач предметной области. В количественной спектроскопии такие свойства связаны характеристиками качества данных: удовлетворение правилам отбора, согласование значений данных в пределах ошибок измерений, согласование порядка следования значений идентифицированных физических величин. Рассмотрены примеры визуализации всех индивидов характеризующих парные отношения между источниками информации.

Наряду с автоматической обработкой данных важным является принятие исследователем решения о качестве данных исходя их просмотра визуализации свойств, характеризующих свойства данных.

Заметим, что в рамках проекта РФФИ частью авторов создаются цифровые библиотеки публикаций по атмосферной химии и радиации [26]. Поскольку эти предметные области содержат значительное число опубликованных данных создаваемые библиотеки создаются с помощью описанной модели публикации, но для решений задач соответствующих предметных областей.

## Литература


[1] MARC 21 Format for Bibliographic Data. http://www.loc.gov/marc/bibliographic

[2] 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.

[3] Functional Requirements for Bibliographic Records, UBCIM Publications – New Series Vol 19, Final Report, 1998. www.ifla.org/files/cataloguing/frbr/frbr.pdf

[4] Лукашевич Н.В., Тезаурусы в задачах информационного поиска, М.: Из-во МГУ, 2011, 512С.

[5] Oberle, D. Semantic management of middleware, Berlin: Springer, 2006. 268 pp.

[6] Privezentsev A., Fazliev A., Tsarkov D., Tennyson J. Computed Knowledge Base for Description of Information Resources of Water Spectroscopy, 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/owled2010_submission_6.pdf

[7] Lavrentiev N.A., Privesentsev A.I., Fazliev A.Z., Filippov N.N. Complete set of published spectral data on $CO_2$ Molecule, Abstracts of the 22-nd Colloquium on High Resolution Molecular Spectroscopy, 2011, p.353.

[8] Voronina S.S., Yurchenko S.N., Fazliev A.Z.Systematization of the published spectroscopic parameters of ammonia, Abstracts of the 22-nd Colloquium on High Resolution Molecular Spectroscopy, 2011, p.163.

[9] Козодоев А.В., Вельмужова И.А., Сенников П.Г., Фазлиев А.З., Филиппов Н.Н., Григорович Н.М. Систематизация опубликованных параметров спектральных линий молекул метана, силана и германа / // Сборник тезисов Международного симпозиума «Атмосферная радиация и динамика» (МСАРД – 2011) Санкт – Петербург, с.102-103.

[10] Половцева Е.Р., Лаврентьев Н.А., Воронина С.С., Науменко О.В., Фазлиев А.З. Информационная система для решения задач молекулярной спектроскопии. 5. Колебательно-вращательные переходы и уровни энергии молекулы $H_2S$, Оптика атм. и океана. 2011, Т.24, №10, с. 898-905.

[11] Лаврентьев Н.А., Привезенцев А.И., Фазлиев А.З. Базы знаний для описания информационных ресурсов в молекулярной спектроскопии 2. Модель данных в количественной спектроскопии, Электронные библиотеки, 2011, т. 14, в.2. http://elbib.ru/index.phtml?page=elbib/rus/journal/2011/part2

[12] Козодоев А.В., Привезенцев А.И. Фазлиев А.З. Аннотирование информационных ресурсов в распределенной информационной системе "Молекулярная спектроскопия", Электронные библиотеки, 2006, т. 9, в.3.

http://www.elbib.ru/index.phtml?page=elbib/rus/journal/2006/ part3/KPF

[13] Быков А.Д., Науменко О.Б., Синица Л.Н., Родимова О.Б., Творогов С.Д., Тонков М.В., Фазлиев А.З., Филиппов Н.Н., Информационные аспекты молекулярной спектроскопии, Томск, Из-во ИОА СО РАН, 2008, 360 с.

[14] Привезенцев А.И., Царьков Д.В., Фазлиев А.З., Базы знаний для описания информационных ресурсов в молекулярной спектроскопии. 3. Формирование базовой и прикладной онтологии, Электронные библиотеки, 2012, т. 15, в.2. http://elbib.ru/index.phtml?page=elbib/rus/journal/2012/part2

[15] Dubernet M.L., Boudon V., Culhane L. et al., Virtual atomic and molecular data centre, J. Quant. Spectrosc. & Rad. Transfer. 2010. v. 111, No 15. p. 2151-2159.

[16] Jacquinet-Husson N., Scott N.A., Chedin A. et al., The GEISA spectroscopic database: current and future archive for earth and planetary atmosphere studies, J. Quant. Spectrosc. & Rad. Transfer. 2008. v. 109. No 6. p. 1043-1059.

[17] Tashkun S.A. and Perevalov V.I. CDSD-4000: High-Temperature Spectroscopic $CO_2$, The 11th HITRAN Database Conference, June 16 – June 18, 2010, Cambridge, 2010, p. 10.

[18] Toth R.A., Brown L.R., Miller C.E., Devi V. Malathy and Benner D.Chris. Spectroscopic database of $CO_2$ line parameters: 4300–7000 $cm^{-1}$, J. Quant. Spectrosc. & Rad. Transfer.2008, v. 109, No 6, p. 906-921.

[19] Rothman L.S., Gordon I.E., Barbe A. et al. The HITRAN 2008 molecular spectroscopic database, J. Quant. Spectrosc. & Rad.Transfer. 2009. v. 110, No 9. p. 533-572.

[20] Ахлёстин А.Ю., Козодоев А.В., Лаврентьев Н.А., Привезенцев А.И., Фазлиев А.З. Базы знаний для описания информационных ресурсов в молекулярной спектроскопии, 4. Программное обеспечение // Электронные библиотеки, 2012, т. 15, в.3. http://elbib.ru/index.phtml?page=elbib/rus/journal/2012/part3/AKLPF

[21] R.A. Toth, C.E. Miller, L.R. Brown, V. Malathy Devi and D. Chris Benner, Line positions and strengths of $^{16}O^{12}C^{18}O$, $^{18}O^{12}C^{18}O$ and $^{17}O^{12}C^{18}O$ between 2200 and 7000 $cm^{-1}$, J. Quant. Spectrosc. & Rad.Transfer., 2007, Volume 243, Issue 1, Pages 43-61

[22] З.В. Апанович, П.С. Винокуров, Т.А. Кислицина. Методы и средства визуализации информационного наполнения больших научных порталов , Вестник НГУ Серия: Информационные технологии. 2011— том 9, выпуск 3, с. 5-14.

[23] A. Noack. Energy Models for Graph Clustering, Journal of Graph Algorithms and Applications, 11(2):453-480, 2007.

[24] Fruchterman T. M. J., Reingold E. M. Graph Drawing by Force-Directed Placement, Software - Practice and Experience, 1991, Vol. 21, N11, P. 1129-1164.

[25] Kamada, T., Kawai, S. An algorithm for drawing general undirected graphs, Information Processing Letters,Vol. 31, 1989, pp. 7-15.

[26] К. М. Фирсов, В.А. Фролькис, Ю. В. Воронина, А.И. Козодоев, А. З. Фазлиев. Распределенная информационная система для атмосферных наук, Материалы 15 Всероссийской конференции «Интернет и современное общество», СПб., 10-12 октября 2012.

## Digital scientific library of quantitative spectroscopy publications

Zinaida Apanovich, Pavel Vinokurov,
Alexey Akhlyostin, Alexey Privezentsev,
Alexander Fazliev

A method of developing a digital library of scientific articles for the domain in which the factological part is significantly bigger than the notional one is being discussed in the report. Using the example of a library of articles on quantitative spectroscopy we demonstrate how the use of a publication (article) model containing domain problems solutions and their properties leads to automated cataloguing of solutions.

Visualization of ontology individuals characterizing the pairs of information sources was of significant importance in this work. Visualization allows one to get a qualitative estimation of consistency of spectroscopy problems' solutions in the case of analyzing huge amount of data.