Различие онтологических представлений предметной области

© С.С. Воронина[1] ©А.И. Привезенцев[1] ©Д.В. Царьков[2] ©А.З. Фазлиев[1]
[1] Институт оптики атмосферы СО РАН,
Томск, Россия
[2] School of Computer Science, the University of Manchester,
Manchester, UK
vss@iao.ru remake@iao.ru tsarkov@cs.man.ac.uk faz@iao.ru

Аннотация

В докладе описаны два онтологических представления предметной области «Количественная спектроскопия». Они включают в себя понятия и термины как абстрактных, так и физических предметных областей. Результатами этих представлений являются онтологии информационных ресурсов количественной спектроскопии и онтологии состояний и процессов количественной спектроскопии. Каждой онтологии соответствует своя группа информационных задач.

Характерной особенностью онтологии информационных ресурсов является значительное число классов и свойств, а онтологии состояний и процессов — число индивидов, превышающее на один-два порядка число классов и свойств.

Оба представления знаний используют коллекцию источников спектральных данных из информационной системы W@DIS. В работе определен круг решаемых информационных задач в каждом из представлений и проведено сравнение метрик этих онтологий.

1 Введение

Известно, что при построении логической теории в предметной области могут использоваться разные концептуализации. Это приводит к появлению логических теорий с разными словарями, а набор таких теорий формирует онтологическое соглашение [1]. Условия совместимости произвольной логической теории для данной предметной области с онтологическим соглашением определяются набором требований [1].

В нашей работе рассмотрена предметная область «Количественная спектроскопия» и построенные для нее две онтологии (логические теории): онтология информационных ресурсов по спектроскопии ряда молекул и онтология состояний и переходов этих молекул. Эти теории использованы для построения онтологического соглашения.

Рассмотрены основные концепты этих теорий и информационная модель количественной спектроскопии. В информационной системе предметная область представлена опубликованными информационными ресурсами. Необходимость создания разных онтологий в количественной спектроскопии связана с разными информационными задачами, решаемыми в прикладных науках с помощью спектральных данных.

В работах [2–5] при построении онтологии информационных ресурсов по количественной спектроскопии основное внимание уделено первичным источникам данных и информации. Следовательно, недостатком модели количественной спектроскопии является неявное описание составных источников данных. В работе [6] сделана попытка дать явное описание составных источников на примере описания состояний и переходов молекул. Появление таких источников данных обусловлено, как собственными задачами спектроскопии (эталонные уровни энергии и переходы, достоверные состояния и переходы), так и потребностью прикладных наук, использующих экспертные данные. Следовательно, необходимо дополнять существующую модель цепи прямых и обратных задач другими задачами, которые не являются первичными задачами спектроскопии, связанными с измерениями и вычислениями параметров спектральных линий.

В работе приведены субъектно-предикатные структуры основных индивидов сравниваемых онтологий и дано описание метрик созданных онтологий.

Труды 16-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» — RCDL-2014, Дубна, Россия, 13–16 октября 2014 г.

2 Информационная модель количественной спектроскопии

В данной работе описание модели предметной области **Quantitative Spectroscopy** представляет собой композицию упрощенных моделей предметных областей **Molecular Spectroscopy, Thermodynamic Conditions, Mathematical Relations, Bibliography** и **Photochemistry**.

В предшествующих работах наиболее детально описана модель **Molecular Spectroscopy**, которая ранее представлялась нами в виде сети прямых и обратных задач [7] или набора состояний и переходов изолированной и неизолированной молекулы [6]. Остальные предметные области охарактеризованы в работах [2, 3], а предметная область **Photochemistry** описана в работе [8].

Описание предметной области можно проводить разными способами. Так сделав акцент на качестве источников данных – приходим к описанию информационных ресурсов, тогда как выделение в качестве объектов исследования состояний и переходов молекулы – приводит к описанию мультимножеств состояний и переходов. Построение онтологии информационных ресурсов опирается на понятие источника данных, причем основную роль играет концепт «первичный источник данных». Анализ составных источников данных проводится с помощью первичных источников. Отметим, что составными источниками могут быть экспертные данные или эталонные уровни энергии или вакуумные волновые числа. Построение онтологии состояний и переходов основано на использовании мультимножеств идентичных состояний и переходов. Такие мультимножества в подавляющем большинстве случаев представляются составными источниками данных.

Опишем задачи спектроскопии, анализируемые при онтологических представлениях количественной спектроскопии, и приведем определения первичных и элементарных источников данных.

Таблица 1

Спектроскопические задачи	Тип источника данных	Определение источника данных
T1, T2, T6, T7	первичный	D1
T3, T5	первичный	D2
T4, E	первичный	D3
T8, T9	составной	D4
T10	составной	D5
T11, T12	составной	D6
T13	составной	D7

D1. Все части опубликованного решения задачи Tn (n = 1, 2, 6, 7), дополненные названием молекулы, библиографической ссылкой и названием метода решения задачи (или ссылкой на описание метода) называются первичным источником данных.

D2. Все части опубликованного решения задачи Tn (n = 3, 5), дополненные названием молекулы, библиографической ссылкой, названием метода решения задачи, термодинамическими условиями и уширяющей молекулой называются первичным источником данных.

D3. Все части опубликованного решения задачи Tn (T4 и E), дополненные названием молекулы, библиографической ссылкой, названием метода решения задачи (или ссылкой на описание метода), физической величиной и термодинамическими условиями называются первичным источником данных.

D6. Все части опубликованных решений задачи Tn (n = 1, 2, 6, 7), дополненные названием молекулы, квантовыми числами описываемого состояния (перехода), библиографическими ссылками, названиями методов решения задачи, называются элементарным источником данных по состоянию (переходу) определяемому квантовыми числами.

D7. Все части опубликованных решений задачи Tn (T3 и T5), дополненные названием молекулы, квантовыми числами описываемого состояния (перехода), библиографическими ссылками, названиями методов решения задачи, одинаковыми термодинамическими условиями и уширяющей молекулой называются элементарным источником данных по состоянию (переходу) определяемому квантовыми числами.

Модель данных количественной спектроскопии, содержащая решения задач T1–T7 и E, расширена введением дополнительных 6 информационных задач. Задачи T8 и T9 состоят в вычислении эталонных (референтных) уровней энергии и вакуумные волновые числа. Задача T10 состоит в построении экспертных наборов данных. Задачи T11 и T12 состоят в выборке полного набора идентичных состояний и переходов изолированной молекулы. Задача T12 состоит в выборке переходов неизолированной молекулы, характеризуемых параметрами спектральных линий. Признаком разделения задач на предметные и информационные является невозможность проведения одного эксперимента для получения представляемых данных.

3 Информационные ресурсы

В этой работе рассмотрены только информационные ресурсы, относящиеся к трем молекулам симметрии C_{2v} и C_s. Статистические данные об этих молекулах приведены в табл. 2. Они включают в себя имеющиеся в ИС W@DIS число источников данных, число переходов, которые описанных при решении прямой (T2) и обратной задачи (T6(a)) и число уникальных переходов (T6(b)) параметры которых измерены. Статистика связана с двумя группами основных концептов количественной спектроскопии, рассмотренных в работе.

Таблица 2. Статистические данные о числе источников данных и переходов по трем молекулам и их изотопологам симметрии C_{2v} и C_s в ИС W@DIS

Молекула	Число источников данных	Число переходов			Молекула	Число источников данных	Число переходов		
		T2	T6(a)	T6(b)			T2	T6(a)	T6(b)
H_2S	39	57	35111	24651	H_2O	154	3153126	179014	64884
$H^{33}SH$	10		335	326	$H^{17}OH$	69	15363234	11844	7513
$H^{34}SH$	12		939	831	$H^{18}OH$	91	20108373	39372	19468
HDS	6		2897	2891	HOD	119	29462133	87309	38989
$HD^{34}S$	2		95	95	$H^{17}OD$	16	2696081	1224	1183
D_2S	6		863	794	$H^{18}OD$	33	2720070	20367	13878
$D_2{}^{34}S$	1		64	64	D_2O	46	2985578	40412	34421
SO_2	48	15721	24466	19674	$D_2{}^{17}O$	4	2503245	505	505
$SO^{17}O$	2	178	97	97	$D_2{}^{18}O$	9	2498276	11500	10131
$SO^{18}O$	7		493	391	HTO	3	111	175	175
$S^{17}O^{18}O$	2		70	70	$^{34}SO_2$	26		14587	14529
$S^{17}O_2$	2		102	100	$^{34}SO^{17}O$	3		123	98
$S^{18}O_2$	10		407	407	$^{34}SO^{18}O$	3		129	129
$^{33}SO_2$	10	669	128	127	$^{34}S^{17}O^{18}O$	1		54	54
$^{33}SO^{18}O$	2		19	19	$^{36}SO_2$	1		29	29
$^{33}S^{17}O_2$	1		15	15					

Концепция первичного источника данных, и, связанного с ним источника информации, применялась нами для описания коллекции опубликованных данных. Первичный источник данных в этом подходе являлся элементарным информационным объектом, свойство которого необходимо описать с целью построения логической теории (онтологии) частей публикаций, содержащих спектральные данные, относящиеся к одной из восьми задач спектроскопии [7]. Созданные онтологии содержали ответы на вопросы: в каких статьях были опубликованы значения физических величин, содержат ли они не достоверные данные, каковы бинарные отношения между источниками данных и т.д. Эта концепция решила задачи описания свойств источника данных для всех восьми задач в модели Molecular Spectroscopy [7], но не позволила дать описание состояний и переходов входящих в первичные источники данных.

Для описания переходов и состояний требуется ввести концепции, описывающие состояние и переход молекулы. Эти концепции существенно отличаются от концепции первичного источника данных. При описании перехода и состояния каждый элементарный информационный объект (источник данных) содержит в себе все опубликованные идентичные состояния или переходы с определенными квантовыми числами. В общем случае элементарные источники данных являются составным источником данных. Созданные онтологии состояний и переходов содержат ответы на вопросы о том, как хорошо согласованы между собой вычисленные и измеренные значения параметров состояния и переходов каждой молекулы в ИС W@DIS.

3.1 Подход к систематизации информационных ресурсов

Для систематизации спектральных данных количественной спектроскопии, применялся подход, в основу которого положено соотнесение данных с решениями задач количественной спектроскопии. В наших ранних работах спектральные данные отождествлялись с решениями одной из восьми задач спектроскопии (T1–T7, E). Семь задач (T1–T7) связаны с расчетами и одна задача (E) с измерениями. Решениями последней задача являются значения измеренных спектральных функций (коэффициент поглощения, сечение поглощения, функция пропускания и т.д.).

В данной работе используется классификация веществ, описанная в [9], и учитывается то, что структурные особенности веществ и их наименования регламентируются многочисленными стандартами. Оригинальной частью работы являются построенные в ИС W@DIS информационные объекты, соответствующие веществам и каналам продуктов фотохимических реакций. В экспертных данных информация о каналах продуктов фотохимических реакций появились в 1999 г. [10]. Мы включили эти каналы в качестве свойств для сечений поглощения, при этом руководствовались следующим принципом: конкретные каналы продуктов реакций добавляются в свойства сечений поглощения в случае, когда они (каналы продуктов) входят в интервал измерений, и/или пороговая длина волны ($\lambda_{threshold}$) канала продуктов близка к внешним границам интервала изменений длин. Наряду с каналами продуктов, в

перечень свойств сечения поглощения включены характеристики каналов продуктов фотохимических реакций.

Элементарные источники данных для переходов формировались виртуально из БД переходов изолированных молекул и участвовали в формировании соответствующих индивидов онтологии переходов изолированной молекулы.

3.2 Особенности онтологического описания информационных ресурсов задач T1–T7 и E

На рис. 1, 2 представлены субъектно-предикатные структуры онтологии информационных ресурсов, относящиеся к задачам T6 и E [7]. Прямоугольники соответствуют индивидам, дуги характеризуют объектные свойства, прямоугольники (так называемые номиналы), не содержащие внутри себя свойств, соответствуют индивидам, представляющим собой общеизвестные термины в предметных областях. Особенность этой онтологии состоит в том, что каждый индивид связан с одной публикацией.

В онтологии информационных ресурсов индивиды, характеризующие индивидуальные свойства источников данных, не зависят от времени. Индивиды, характеризующие бинарные отношения между источниками данных, зависят от времени, т.е. они могут быть дополнены свойствами, появление которых связано с занесением источников данных, имеющих идентичные переходы или состояния тем, которые собраны в базе данных ИС W@DIS [11].

3.3 Особенности онтологического описания состояний и переходов

На рис. 3 показаны упрощенные структуры индивидов онтологии состояний и переходов молекул и их связи с физическими величинами, характеризующие параметры спектральных линий, а также предметные области. На рис. 4 в деталях показана структура индивида, описывающего переход 010 726 – 000 717 в изолированной молекуле $H^{17}OD$.

Отдельные части индивида, характеризующего конкретный переход в онтологии переходов, меняются со временем. Причиной является появление новых данных в свежих публикациях. Не меняющейся частью индивида перехода являются индивиды, описывающие квантовые числа.

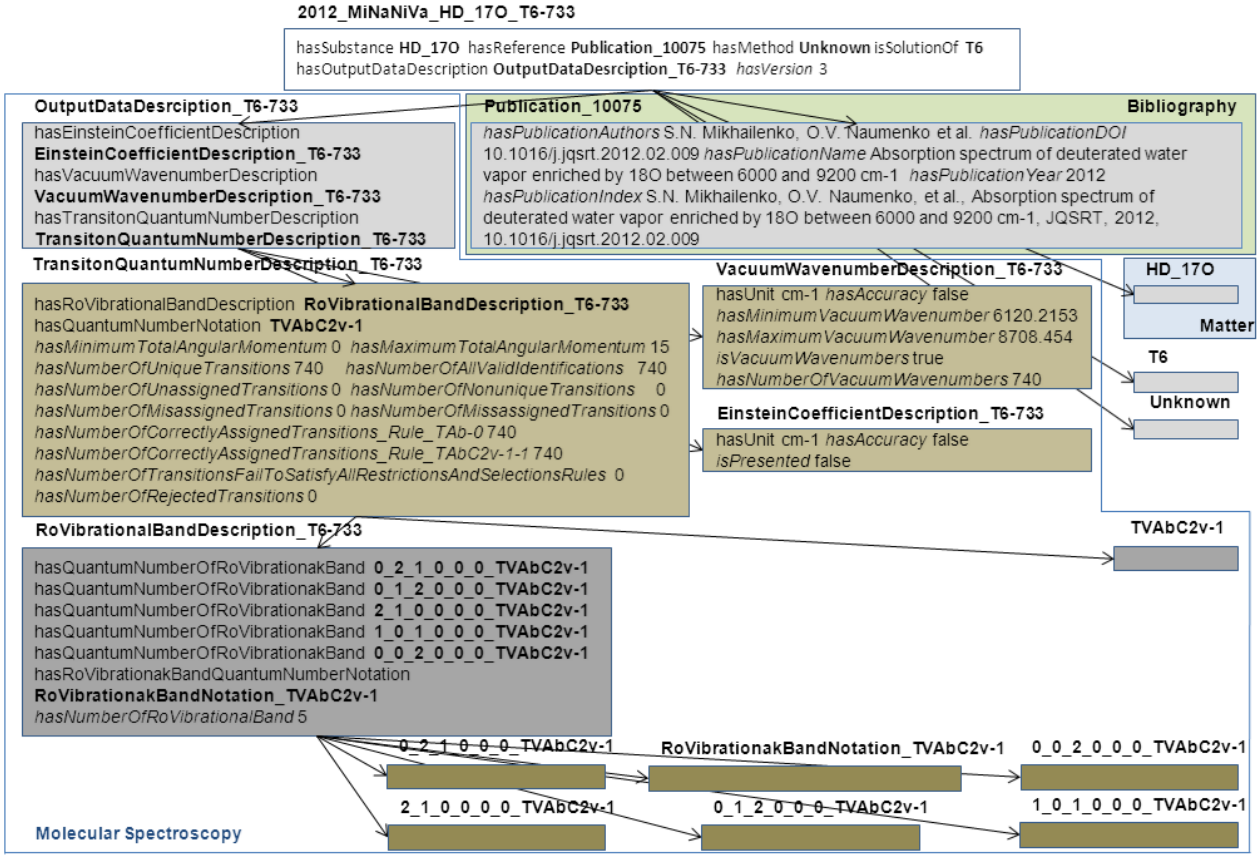

Рис. 1. Субъектно-предикатная структура основного индивида онтологии информационных ресурсов по переходам изолированной молекулы в количественной спектроскопии

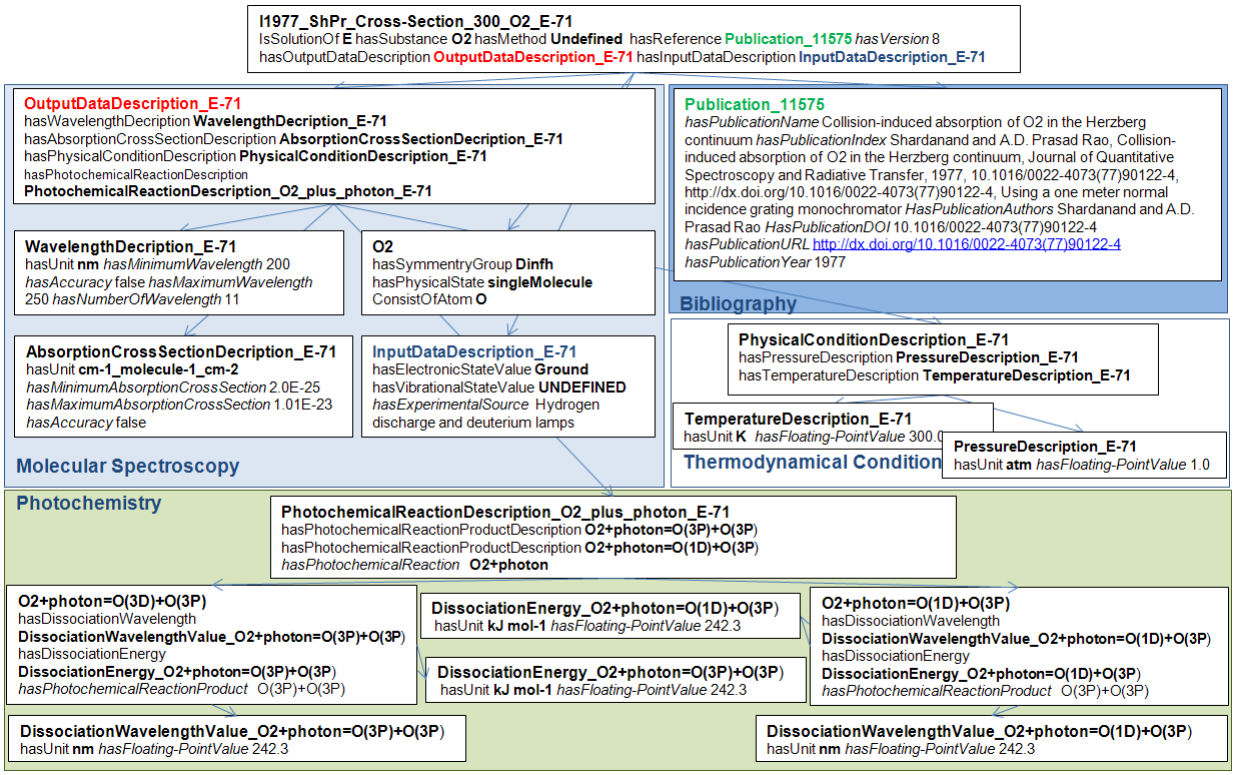

Рис. 2. Субъектно-предикатная структура основного индивида онтологии информационных ресурсов по сечениям поглощения молекулы в количественной спектроскопии

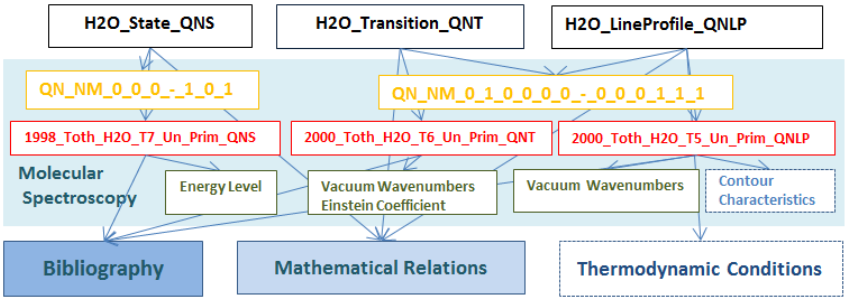

Рис. 3. Структура индивидов предметной области «**Количественная спектроскопия**», входящих в онтологию состояний и переходов молекул

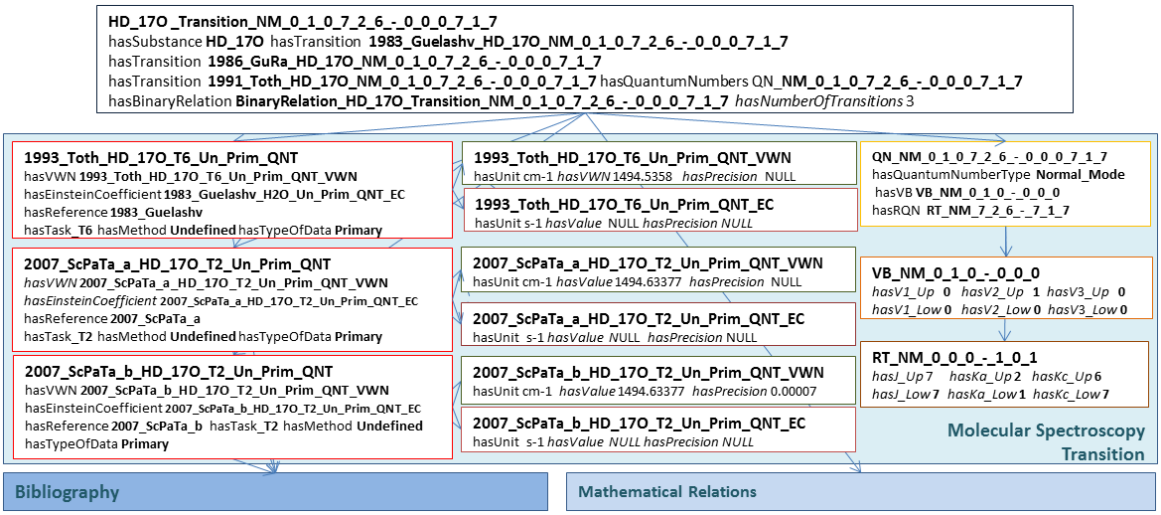

Рис. 4. Структура индивида **HD_17O_Transition_QNT**, QNT = 0_1_0_7_2_6_-_0_0_0_7_1_7, **T6_Un_Prim** = hasTask **T6** – hasMethod **Undefined** – hasTypeOfData **Primary**, **VWN** = vacuum wavenumbers, **EC** – Einstein coefficient

3.5 Сравнение метрик онтологий

В базе данных информационных ресурсов размещено 16 источников данных о переходах изолированной молекулы $HD^{17}O$. Онтологическое описание информационных ресурсов этой молекулы содержит 13778 логических аксиом, 106 классов, 40 объектных и 159 конкретных (datatype) свойств и 4814 индивидов. Онтология переходов этой молекулы и других молекул имеет DL-выразительность ALCHOIN(D).

В ИС W@DIS в базе данных спектральных функций (сечений поглощения) молекул практически не содержатся данные об изотопологах молекул. По этой причине ниже приведены данные по основному изотопологу молекулы воды. В БД содержится 37 источников данных о сечениях поглощения H_2O. Онтология источников информации о сечениях поглощения воды построена на основе 37 источников данных. Она включает 2822 логических аксиом, 101 класс, 40 объектных и 154 конкретных свойств и 620 индивидов. DL выразительность этой онтологии ALCHON (D).

В настоящее время в базе данных переходов для изолированной молекулы $HD^{17}O$ содержится 1183 перехода. Онтологическое описание этой молекулы содержит 93457 логических аксиом, 29 классов, 15 объектных и 25 конкретных (datatype) свойств и 19068 индивидов. Онтология переходов этой молекулы и других молекул имеет DL-выразительность ALCRIF(D). В онтологии информационных ресурсов большинство аксиом (30258) являются аксиомами A-box типа. В отличие от онтологии информационных ресурсов, онтология переходов не содержит номиналов, т.е. все индивиды относятся к A-box.

Заметим, что приведенные примеры для молекулы $HD^{17}O$ указывают на то, что отношение числа индивидов в онтологии переходов и информационных ресурсов приблизительно равно четырем. В хорошо изученных молекулах (например, основных изотопологов воды и диоксида углерода) это число увеличивается на два порядка.

4 Заключение

В работе сделан шаг к построению онтологического соглашения объединяющего логические теории количественной спектроскопии, построенные на основе разных словарей. Рассмотрена модель количественной спектроскопии, обобщающая предыдущие работы авторов. При обобщении основное внимание уделено интерпретации концепта «составной источник данных». С этой целью рассмотрены шесть информационных задач относящиеся к описанию эталонных и экспертных данных и состояний и переходов в количественной спектроскопии. Основное внимание сконцентрировано на онтологиях информационных ресурсов количественной спектроскопии и онтологии состояний и переходов в количественной спектроскопии. Приведены субъектно-предикатные структуры некоторых основных индивидов этих онтологий и описаны метрики соответствующих онтологий.

Литература

[1] D. Oberle, Semantic management of middleware. – Springer, 2006, 268 p.

[2] Привезенцев А. И. Организация онтологических баз и программное обеспечение для описания информационных ресурсов в молекулярной спектроскопии: дис. … канд. техн. наук. – Томск, 2009.

[3] Базы знаний для описания информационных ресурсов в молекулярной спектроскопии 2. Модель данных в количественной спектроскопии // Электронные библиотеки. 2011. Т. 14, вып. 2. http://elbib.ru/index.phtml?page=elbib/rus/journal/2011/part2/LPF

[4] 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, № 2. http://elbib.ru/index.phtml?page=elbib/rus/journal/2012/part2/PTF

[5] Ахлёстин А.Ю., Лаврентьев Н.А., Привезенцев А.И., Фазлиев А.З. Базы знаний для описания информационных ресурсов в молекулярной спектроскопии. 5. Качество экспертных данных // Электронные библиотеки. 2013. Т. 16, № 4. http://www.elbib.ru/index.phtml?page=elbib/rus/journal/2013/part4/AKLPF

[6] S.S. Voronina, A.I. Privezentsev, D.V. Tsarkov and A.Z. Fazliev, An ontological description of states and transitions in quantitative spectroscopy, Proc. of SPIE, 2014 (в печати)

[7] A.D. Bykov, A.V. Kozodoev, A.I. Privezentsev, L.N.Sinitsa, M.V.Tonkov, A.Z.Faliev, N.N.Filippov, M.Yu. Tretyakov, Distributed information system on molecular spectroscopy, Proc. of SPIE, International Symposium on High Resolution Molecular Spectroscopy, 2006, vol. 6580, pp. 65800W

[8] Yu. Voronina, N. Lavrentiev, V. Privezentsev, A. Fazliev, and K. Firsov, Representation of absorption cross-sections in information system W@DIS, Proc. of SPIE, 2014 (в печати)

[9] H. Keller-Rudek, G.K. Moortgat, R. Sander, and R. Sorensen, The MPI-Mainz UV/VIS Spectral Atlas of Gaseous Molecules of Atmospheric Interest, Earth Syst. Sci. Data, 5, 365–373, 2013, doi:10.5194/essd-5-365

[10] R. Atkinson, D.L. Baulch, R.A. Cox, R.F. Hampson, J.A. Kerr, M.J. Rossi, and J. Troe, Evaluated Kinetic and Photochemical Data for Atmospheric Chemistry, Organic Species: Supplement VII, J. Phys. Chem. Ref. Data 28(2), 191 (1999).

[11] ИС W@DIS. http://wadis.saga.iao.ru

[12] L.S. Rothman, I.E. Gordon, Y. Babikov, et al., The HITRAN 2012 Molecular Spectroscopic Database, JQSRT, 2013, Vol. 130, P. 4–50, DOI: 10.1016/j.jqsrt.2013.07.002.

[13] N. Jacquinet-Husson, L. Crepeau, R. Armante, et al., The 2009 edition of the GEISA spectroscopic database, JQSRT, 2011, Vol. 112, Iss. 15, P. 2395–2445, DOI: 10.1016/j.jqsrt.2011.06.004.

[14] M.L. Dubernet, V. Boudon, J.L. Culhane, et al., Virtual atomic and molecular data centre, JQSRT, 2010, Vol. 111, Iss. 15, P. 2151–2159, DOI: 10.1016/j.jqsrt.2010.05.004.

Clear-cut Distinction between Domain Ontological Representations

S.S. Voronina, A.I. Privezentsev, D.V. Tsarkov, A.Z. Fazliev

A paper contains a brief description of two ontological representations of quantitative spectroscopy. These representations are based on concepts and relations related to some abstract and physical domains. Ontology of information resources of quantitative spectroscopy and ontology of states and transitions of molecules are the results of these knowledge representations. First ontology is focused on abstract properties of the information sources related to quantitative spectroscopy, while the second one is focused on a description of physical behavior of a molecule.

A considerable number of classes and properties is a specific feature of ontology of information resources, so the considerable number of individuals characterizes the ontology of states and transitions of molecule.

The collection of spectral data sources from information system W@DIS was used to demonstrate the differences between these knowledge representations. A final part of the paper contains a comparison of metrics of the ontologies.