«Оптика атмосферы и океана», **35**, № 3 (2022)
DOI: 10.15372/AOO20220302

УДК 539.194

Интеграция параметров спектральных линий молекул CO_2, N_2O, NO_2, и C_2H_2 в узле LTS распределенной информационной системы VAMDC

Р.В. Кочанов*

Институт оптики атмосферы им. В.Е. Зуева СО РАН
634055, г. Томск, пл. Академика Зуева, 1

Поступила в редакцию 25.10.2021 г.

Рассматривается интеграция параметров спектральных линий банков данных CDSD, NOSD, NDSD и ASD, разработанных в Лаборатории теоретической спектроскопии ИОА СО РАН, в единый узел LTS Виртуального центра атомных и молекулярных данных (VAMDC). Написано программное обеспечение, реализующее обмен данными в формате XML с порталом VAMDC, а также интеграцию списков линий больших объемов в базу данных.

Ключевые слова: VAMDC, банк данных, распределенная система, CDSD-296, NDSD-1000, NOSD-1000, ASD-1000; VAMDC, data bank, database, distributed system, CDSD-296, NDSD-1000, NOSD-1000, ASD-1000.

Введение

Вопрос стандартизации структур данных актуален как в области спектроскопии, так и в других научных областях. Основная трудность стандартизации спектроскопических данных связана с гетерогенностью их структур, обусловленной, в частности, различной симметрией молекул. Публикация данных разного формата в статьях также затрудняет импорт данных, их интеграцию и сравнение. Задача стандартизации структур спектральных данных решена в международном проекте Виртуального центра атомных и молекулярных данных VAMDC (The Virtual Atomic and Molecular Data Centre) [1, 2], и ее решение представлено в форме XML-схемы XSAMS. VAMDC объединяет более тридцати исследовательских групп и включает в себя данные по спектрам атомов и молекул для широкого круга приложений. К числу баз данных (БД), интегрированных в распределенную информационную систему (РИС) VAMDC, относятся БД информационных систем и банков данных HITRAN [3], CDMS [4], S&MPO [5], W@DIS [6], EXOMOL [7] и др.

Упрощенная схема распределенной инфраструктуры VAMDC представлена на рис. 1. Узлы обмениваются информацией с пользователем по интернет-протоколу VAMDC-TAP (https://standards.vamdc.eu); они могут работать как через веб-портал (https://portal.vamdc.org), так и в режиме отдельных серверов данных для стороннего ПО.

Основой для типизации структур спектральных данных является XML-схема XSAMS (XML Schema for Atoms, Molecules and Solids) [1, 8]. Она позволяет структурировать спектральные данные и проверять передаваемые данные об атомах, молекулах и твердых телах между узлом и сервером VAMDC. Следует отметить, что схема и XML-документы, используемые при передаче данных, не предназначены для того, чтобы с ними работали конечные пользователи. Созданные XML-документы, использующие XSAMS-схему, должны впоследствии быть преобразованы для более наглядного и удобного представления данных (таблицы, рисунки, диаграммы и т.д.) с помощью специальных программ-«конверторов».

Постановка задачи

В Институте оптики атмосферы (ИОА) СО РАН разрабатываются банки данных по спектральным переходам многоатомных молекул, представляющих интерес для атмосферных и астрофизических приложений. При ключевом участии Лаборатории теоретической спектроскопии (ЛТС) ИОА СО РАН были разработаны и опубликованы банки данных для молекул CO_2 (CDSD-296 [9], CDSD-1000 [10], CDSD-4000 [11]), N_2O (NOSD [12]), NO_2 (NDSD [13]) и C_2H_2 (ASD [14]). В 2014 г. банки данных CDSD были интегрированы в VAMDC в качестве трех отдельных узлов [15]. Позднее были созданы и интегрированы в РИС VAMDC узлы для банков данных NOSD, NDSD и ASD [1].

На момент первичной интеграции банков данных ЛТС в VAMDC (2011 г.) в силу того, что структуры спектральных данных для каждого из этих банков не совпадали, создание единого узла LTS в РИС VAMDC (далее узел LTS), вмещающего все шесть банков, было затруднено. Поэтому было принято решение создавать отдельные узлы для разных групп молекул. Каждый узел содержал базу спек-

* Роман Викторович Кочанов (roman2400@rambler.ru).

© Кочанов Р.В., 2022

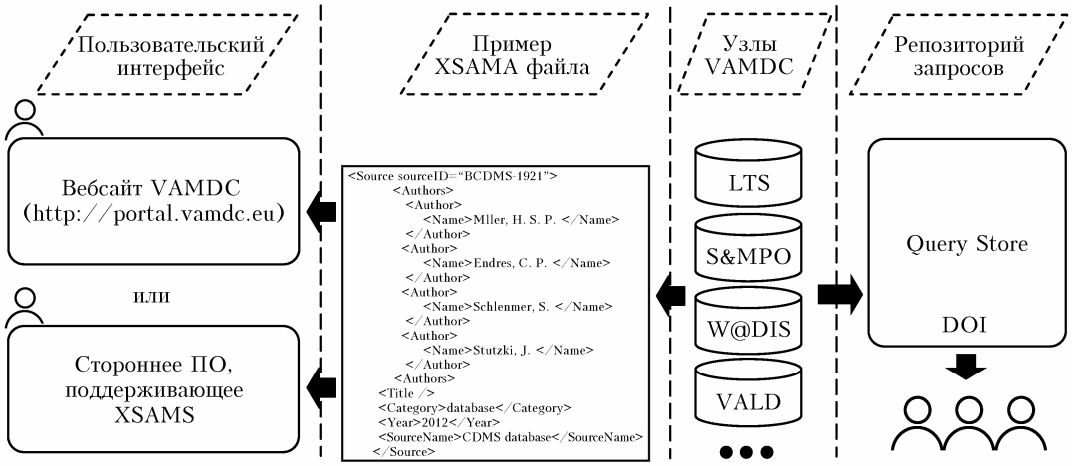

Рис. 1. Общая схема инфраструктуры VAMDC. Узлы системы приведены в качестве примера. Полный список актуальных узлов представлен на сайте VAMDC (https://vamdc.org)

тральных данных, относящуюся к молекулам одной симметрии. В определенный момент поддерживать и типизировать новые узлы стало обременительно.

По сути, объединение банков данных подразумевает создание базы данных с таблицами для каждой группы молекул одной симметрии. Строки такой таблицы содержат параметры спектральных линий, и каждая строка соответствует данным из одной определенной публикации.

В настоящей работе рассматривается создание объединенного узла LTS, включающего в себя спектроскопические банки данных для многоатомных молекул, содержащие результаты расчетов в ИОА СО РАН. Текущая версия узла, о которой идет речь в работе, включает банки данных CDSD-296, NOSD-1000, NDSD-1000 и ASD-1000.

Банки данных

В табл. 1 дана сводка основных параметров в четырех банках данных, добавленных в единый узел LTS. Версия CDSD-296, которая была включена в узел, содержит недавние исправления (ftp://ftp.iao.ru/pub/CDSD-296/, readme_version_1.txt).

Таблица 1

Статистика основных параметров в банках данных, входящих в состав нового узла LTS VAMDC. Источники данных: по CO_2 – CDSD-296 [9], NO_2 – NDSD-1000 [13], N_2O – NOSD-1000 [12], C_2H_2 – ASD-1000 [14]

ID	Изотополог	N_{lines}	ν_{min}, см$^{-1}$	ν_{max}, см$^{-1}$	S_{min}, см$^{-1}$/(мол. · см$^{-2}$)	S_{max}, см$^{-1}$/(мол. · см$^{-2}$)
			CO_2			
1	$(^{12}C)(^{16}O)_2$	170388	345,937	14075,301	1,0e-30	3,55e-18
2	$(^{13}C)(^{16}O)_2$	70226	408,380	13734,963	1,0e-30	3,74e-20
3	$(^{16}O)(^{12}C)(^{18}O)$	115966	410,919	12677,182	1,0e-30	6,80e-21
4	$(^{16}O)(^{12}C)(^{17}O)$	71992	429,164	12726,562	1,0e-30	1,26e-21
5	$(^{16}O)(^{13}C)(^{18}O)$	40682	461,995	9212,608	1,0e-30	7,96e-23
6	$(^{16}O)(^{13}C)(^{17}O)$	22932	493,882	8061,739	1,0e-30	1,39e-23
7	$(^{12}C)(^{18}O)_2$	10591	482,814	8162,828	1,0e-30	1,32e-23
8	$(^{17}O)(^{12}C)(^{18}O)$	14746	498,617	6908,462	1,0e-30	2,53e-24
9	$(^{12}C)(^{17}O)_2$	6599	535,357	6932,954	1,0e-30	2,88e-25
10	$(^{13}C)(^{18}O)_2$	3112	539,620	6687,643	1,0e-30	1,41e-25
11	$(^{17}O)(^{13}C)(^{18}O)$	3422	555,754	3602,575	1,0e-30	2,71e-26
12	$(^{13}C)(^{17}O)_2$	1657	575,853	3616,620	1,0e-30	3,01e-27
	TOTAL CO_2	**532313**	**345,936**	**14075,301**	**1,0e-30**	**3,55e-18**
			NO_2			
1	$(^{14}N)(^{16}O)_2$	1046808	466,611	4775,32	2,29e-47	1,28e-19
			N_2O			
1	$(^{14}N)_2(^{16}O)$	1405069	3,352	8726,404	1,0e-25	1,55e-19
			C_2H_2			
1	$(^{12}C)_2H_2$	33890981	2,554	10002,587	4,32E-49	1,20e-18

Примечание. N_{lines} – количество линий; ν_{min}, ν_{max} – минимальное и максимальное значение волнового числа; S_{min}, S_{max} – минимальное и максимальное значение интенсивности. Нумерация изотопологов соответствует номенклатуре HITRAN (https://hitran.org/docs/iso-meta).

Единый узел LTS РИС VAMDC

Импорт банков данных (параметров спектральных линий молекул разной симметрии) в одну базу данных позволит значительно облегчить поддержку узла LTS в РИС VAMDC, упростить его программное обеспечение и процесс добавления новых изотопологов и молекул. По сути, объединение спектроскопических банков подразумевает создание единой БД по молекулам, изотопологам, спектральным переходам и литературным источникам. Для этой цели было написано программное обеспечение, реализующее обмен данными с порталом VAMDC (подготовка XML-документа для передачи в ответ на запрос портала VAMDC), спроектирована реляционная схема данных, а также создан программный инструментарий по добавлению новых данных в узел.

В силу специфики рассматриваемых банков данных (большое количество спектральных линий, вплоть до десятков миллионов для некоторых банков) единая база данных должна удовлетворять требованиям высокого быстродействия и целостности данных. В качестве системы управления базами данных (СУБД) была выбрана система Yandex Clickhouse (https://clickhouse.com). Это альтернатива более распространенным реляционным СУБД (MySQL https://mysql.com, PostgreSQL https://postgresql.org и др.), предназначенная для хранения и анализа больших массивов спектральных линий. В рамках этой СУБД целостность данных обеспечивается строгой типизацией атрибутов отношений и использованием первичных ключей.

Спроектированная схема реляционной базы данных показана на рис. 2. Для изотопологов и молекул, помимо основных параметров, в БД заносятся идентификаторы IUPAC (InChI и InChIKey), а также идентификатор HITRAN (AFGL). Состояния для изотопологов с разными свойствами симметрии обозначаются разными наборами квантовых чисел.

В табл. 2 представлен список параметров спектральных линий, значения которых вошли в единый узел VAMDC. Столбец «Обозначение в XSAMS» содержит путь в схеме данных XSAMS до соответствующего параметра. Полный набор параметров разделен на две части (см. официальную документацию на сайте https://vamdc.org): первый раздел «*Processes.Radiative.RadiativeTransition*» содержит параметры спектроскопических переходов, второй раздел «*Species.Molecules.Molecule.MolecularState.MolecularStateCharacterisation*» — параметры состояний, такие как квантовые числа, энергии и статистические веса.

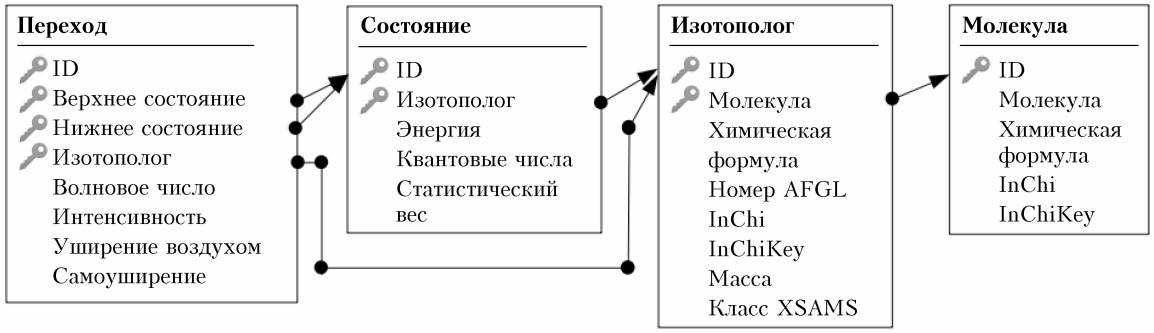

Рис. 2. Схема реляционной БД для узла LTS. Символ ключа на схеме означает специальные столбцы таблицы — так называемые «первичные» и «внешние» ключи. Наличие этих столбцов обязательно для того, чтобы обеспечить ссылочную целостность базы данных

Таблица 2

Параметры спектральных линий, значения которых содержатся в едином узле LTS

Параметр	Наличие в БД	Описание	Обозначение в XSAMS
1	2	3	4
Раздел XSAMS «Processes.Radiative.RadiativeTransition»			
ν	++++	Волновое число перехода	EnergyWavelength.Wavenumber
S	++++	Интенсивность перехода	Probability.LineStrength
A	++++	Коэффициент Эйнштейна	Probability.TransitionProbabilityA
γ_{air}	++++	Коэффициент уширения воздухом	Lineshape.LineshapeParameter
n_{air}	++++	Показатель температурной зависимости уширения воздухом	Lineshape.LineshapeParameter
δ_{air}	+–+–	Коэффициент температурной зависимости сдвига воздухом	LineshapeParameter
γ_{self}	++++	Коэффициент самоуширения	Lineshape.LineshapeParameter
n_{self}	+++–	Показатель температурной зависимости самоуширения	LineshapeParameter
δ_{self}	+–+–	Коэффициент температурной зависимости самосдвига	LineshapeParameter

Окончание табл. 2

1	2	3	4
Раздел XSAMS «Species.Molecules.Molecule.MolecularState.MolecularStateCharacterisation»			
E''	++++	Энергия нижнего состояния	StateEnergy
VIB′	++++	Колебательная идентификация верхнего состояния	Case
VIB″	++++	Колебательная идентификация нижнего состояния	Case
ROT′	++++	Вращательная идентификация верхнего состояния	Case
ROT″	++++	Вращательная идентификация нижнего состояния	Case
g'	+−−+	Статистический вес верхнего состояния	TotalStatisticalWeight
g''	+−−+	Статистический вес нижнего состояния	TotalStatisticalWeight

Примечание. + или − означает, содержится ли этот параметр в исходных версиях банков данных CDSD, NDSD, NOSD и ASD соответственно.

Поскольку архитектура узла и структура БД рассчитаны на эффективную обработку больших данных (в перспективе сотни миллионов спектральных линий и больше), к основным таблицам, содержащим параметры спектральных линий, была применена частичная денормализация. Это подразумевает сохранение части информации о квантовых числах состояний в строках таблицы, содержащей квантовые переходы. Такой прием позволяет добиться меньшей задержки при запросах больших объемов спектральных переходов и контроля соответствия структуре данных по XSAMS-схеме. Также, поскольку формат данных в узле LTS в целом отвечает стандартам HITRAN, некоторые метаданные не заносятся в базу в явном виде, например данные о референсных температурах для параметров контуров линий. В будущем планируется расширение базы данных для добавления более широкого ряда сущностей (статистические суммы, контуры линий, потенциальные функции и т.д.).

Заключение

В настоящей работе рассмотрено объединение узлов РИС VAMDC, содержащих банки данных CDSD-296 [9], NOSD-1000 [12], NDSD-1000 [13], ASD-1000 [14], поддерживаемых ИОА СО РАН, в единый узел LTS. Для объединения создано программное обеспечение, призванное упростить добавление в узел новых банков данных большого объема за счет новой архитектуры реляционной БД, позволяющей включать списки спектральных линий молекул разной симметрии. Быстрое выполнение запросов и эффективная обработка больших массивов данных спектральных линий возможна благодаря СУБД Yandex Clickhouse.

Работа выполнена в рамках государственного задания ИОА СО РАН.

1. *Albert D., Antony B.K., Ba Y.A., Babikov Y.L., Bollard P., Boudon V., Delahaye F., Del Zanna G., Dimitrijević M.S., Drouin B.J., Dubernet M.-L., Duensing F., Emoto M., Endres C.P., Fazliev A.Z., Glorian J.-M., Gordon I.E., Gratier P., Hill C., Jevremović D., Joblin C., Kwon D.-H., Kochanov R.V., Krishnakumar E., Leto G., Loboda P.A., Lukashevskaya A.A., Lyulin O.M., Marinković B.P., Markwick A., Marquart T., Mason N.J., Mendoza C., Millar T.J., Moreau N., Morozov S.V., Möller T., Müller H.S.P., Mulas G., Murakami I., Pakhomov Y., Palmeri P., Penguen J., Perevalov V.I., Piskunov N., Postler J., Privezentsev A.I., Quinet P., Ralchenko Y., Rhee Y.-J., Richard C., Rixon G., Rothman L.S., Roueff E., Ryabchikova T., Sahal-Bréchot S., Scheier P., Schilke P., Schlemmer S., Smith K.W., Schmitt B., Skobelev I.Yu., Srecković V.A., Stempels E., Tashkun S.A., Tennyson J., Tyuterev V.G., Vastel Ch., Vujčić V., Wakelam V., Walton N.A., Zeippen C., Zwölf C.M.* A decade with VAMDC: Results and ambitions // Atoms. 2020. V. 8. P. 76. DOI: 10.3390/atoms8040076.
2. *Dubernet M.L., Antony B.K., Ba Y.A., Babikov Yu.L., Bartschat K., Boudon V., Braams B.J., Chung H.-K., Daniel F., Delahaye F., Del Zanna G., de Urquijo J., Dimitrijević M.S., Domaracka A., Doronin M., Drouin B.J., Endres C.P., Fazliev A.Z., Gagarin S.V., Gordon I.E., Gratier P., Heiter U., Hill C., Jevremović D., Joblin C., Kasprzak A., Krishnakumar E., Leto G., Loboda P.A., Louge T., Maclot S., Marinković B.P., Markwick A., Marquart T., Mason H.E., Mason N.J., Mendoza C., Mihajlov A.A., Millar T.J., Moreau N., Mulas G., Pakhomov Yu., Palmeri P., Pancheshnyi S., Perevalov V.I., Piskunov N., Postler J., Quinet P., Quintas-Sánchez E., Ralchenko Yu., Rhee Y.-J., Rixon G., Rothman L.S., Roueff E., Ryabchikova T., Sahal-Bréchot S., Scheier P., Schlemmer S., Schmitt B., Stempels E., Tashkun S., Tennyson J., Tyuterev Vl.G., Vujčić V., Wakelam V., Walton N.A., Zatsarinny O., Zeippen C.J., Zwölf C.M.* The virtual atomic and molecular data centre (VAMDC) consortium // J. Phys. B. At. Mol. Opt. Phys. 2016. V. 49. P. 074003.
3. *Gordon I.E., Rothman L.S., Hargreaves R.J., Hashemi R., Karlovets E.V., Skinner F.M., Conway E.K. Hill C., Kochanov R.V., Tan Y., Wcisło P., Finenko A.A., Nelson K., Bernath P.F., Birk M., Boudon V., Campargue A., Chance K.V., Coustenis A., Drouin B.J., Flaud J.-M., Gamache R.R., Hodges J.T., Jacquemart D., Mlawer E.J., Nikitin A.V., Perevalov V.I., Rotger M., Tennyson J., Toon G.C., Tran H., Tyuterev V.G., Adkins E.M., Baker A., Barbe A., Canu E., Császár A.G., Dudaryonok A., Egorov O., Fleisher A.J., Fleurbaey H., Foltynowicz A., Furtenbacher T., Harrison J.J., Hartmann J.-M., Horneman V.-M., Huang X., Karman T., Karns J., Kassi S., Kleiner I., Kofman V., Kwabia–Tchana F., Lavrentieva N.N., Lee T.J., Long D.A., Lukashevskaya A.A., Lyulin O.M., Makhnev V.Yu., Matt W., Massie S.T., Melosso M., Mikhailenko S.N., Mondelain D., Müller H.S.P., Naumenko O.V., Perrin A., Polyansky O.L., Raddaoui E.,*

Raston P.L., Reed Z.D., Rey M., Richard C., Tybiás R., Sadiek I., Schwenke D.W., Starikova E., Sung K., Tamassia F., Tashkun S.A., Vander Auwera J., Vasilenko I.A., Vigasin A.A., Villanueva G.L., Vispoel B., Wagner G., Yachmenev A., Yurchenko S.N. The HITRAN2020 molecular spectroscopic database // J. Quant. Spectrosc. Radiat. Transfer. 2022. V. 277. DOI: 10.1016/j.jqsrt.2021.107949.

4. *Endres C.P., Schlemmer S., Schilke P., Stutzki J., Müller H.S.P.* The Cologne Database for Molecular Spectroscopy, CDMS, in the Virtual Atomic and Molecular Data Centre, VAMDC // J. Mol. Spectrosc. 2016. V. 327. P. 95–104. DOI:10.1016/j.jms.2016.03.005.
5. *Babikov Y.L., Mikhailenko S.N., Barbe A., Tyuterev V.G.* S&MPO – an information system for ozone spectroscopy on the WEB // J. Quant. Spectrosc. Radiat. Transfer. 2014. V. 145. P. 169–196. DOI: 10.1016/j.jqsrt.2014.04.024.
6. *Akhlyostin A., Apanovich Z., Fazliev A., Kozodoev A., Lavrentiev N., Privezentsev A., Rodimova O., Voronina S., Császár A.G., Tennyson J.* The current status of the W@DIS information system // Proc. SPIE. 2016. V. 10035. P. 100350D. DOI: 10.1117/12.2249235.
7. *Tennyson J., Yurchenko S.N., Al-Refaie A.F., Barton E.J., Chubb K.L., Coles P.A., Diamantopoulou S., Gorman M.N., Hill C., Lam A.X., Lodi L., McKemmish L.K., Na Yu., Owens A., Polyansky O.L., Rivlin T., Sousa-Silva C., Underwood D.S., Yachmenev A., Zak E.* The ExoMol database: Molecular line lists for exoplanet and other hot atmospheres // J. Mol. Spectrosc. 2016. V. 327. P. 73–94. DOI: 10.1016/j.jms.2016.05.002.
8. *Zwolf C.M., Moreau N., Dubernet M.L.* New model for datasets citation and extraction reproducibility in VAMDC // J. Mol. Spectrosc. 2016. V. 327. P. 122–137. DOI: 10.1016/j.jms.2016.04.009.
9. *Tashkun S.A., Perevalov V.I., Gamache R.R., Lamouroux J.* CDSD-296, high-resolution carbon dioxide spectroscopic databank: An update // J. Quant. Spectrosc. Radiat. Transfer. 2019. V. 228. P. 124–131. DOI: 10.1016/j.jqsrt.2019.03.001.
10. *Tashkun S.A., Perevalov V.I., Teffo J.-L., Bykov A.D., Lavrentieva N.N.* CDSD-1000, the high-temperature carbon dioxide spectroscopic databank // J. Quant. Spectrosc. Radiat. Transfer. 2003. V. 82. P. 165–196. DOI: 10.1016/S0022-4073(03)00152-3.
11. *Tashkun S.A., Perevalov V.I.* CDSD-4000: High-resolution, high-temperature carbon dioxide spectroscopic databank // J. Quant. Spectrosc. Radiat. Transfer. 2011. V. 112. P. 1403–1410. DOI: 10.1016/j.jqsrt.2011.03.005.
12. *Tashkun S.A., Perevalov V.I., Lavrentieva N.N.* NOSD-1000, the high-temperature nitrous oxide spectroscopic databank // J. Quant. Spectrosc. Radiat. Transf. 2016. V. 177. DOI: 10.1016/j.jqsrt.2015.11.014.
13. *Lukashevskaya A.A., Lavrentieva N.N., Dudaryonok A.C., Perevalov V.I.* NDSD-1000: High-resolution, high-temperature Nitrogen Dioxide Spectroscopic Databank // J. Quant. Spectrosc. Radiat. Transfer. 2016. V. 184. P. 205–217. DOI: 10.1016/j.jqsrt.2016.07.014.
14. *Lyulin O.M., Perevalov V.I.* ASD-1000: High-resolution, high-temperature acetylene spectroscopic databank // J. Quant. Spectrosc. Radiat. Transfer. 2017. V. 201. P. 94–103. DOI: 10.1016/j.jqsrt.2017.06.032.
15. *Кочанов Р.В., Перевалов В.И., Ташкун С.А.* Интеграция параметров спектральных линий молекулы CO_2, содержащихся в банках данных CDSD, в Виртуальный центр атомных и молекулярных данных (VAMDC) // Оптика атмосф. и океана. 2014. Т. 27, № 3. С. 240–245; *Kochanov R.V., Perevalov V.I., Tashkun S.A.* Integration of CO_2 spectral line parameters from the CDSD databanks into the virtual atomic and molecular data center (VAMDC) // Atmos. Ocean. Opt. 2014. V. 27. DOI: 10.1134/S1024856014060098.

R.V. Kochanov. **Integration of the spectral line parameters of CO_2, N_2O, NO_2, and C_2H_2 into the node of the distributed information system VAMDC.**

The integration of the parameters of the CDSD, NOSD, NDSD, and ASD data banks developed at the Laboratory of Theoretical Spectroscopy, Institute of Atmospheric Optics, Siberian Branch, Russian Academy of Sciences, into a single node of the Virtual Atomic and Molecular Data Center (VAMDC) is described. As part of this work, software was written that implements XML-formatted data exchange with the VAMDC portal and integrates the data banks into the distributed infrastructure.