Improvement of the Data Quality Assessment Procedure in Large Collections of Spectral Data

Alexey Akhlestin, Nikolai Lavrentiev, Alexey Kozodoev,

Elena Kozodoeva, Alexey Privezentsev, Alexander Fazliev[0000-0003-2625-3156]

V. E. Zuev Institute of Atmospheric Optics SB RAS, Tomsk, Russia
lexa@iao.ru, lnick@iao.ru, kav@iao.ru, faz@iao.ru,
remake@iao.ru, klen@iao.ru

Abstract. Relations between the components of the data layer and application layer in the quantitative spectroscopy are schematically represented. The sequence of actions in the data quality analysis in the W@DIS information system is described. Methods for the spectral data quality analysis and assessment of trust in expert data sources and the results of their application in the W@DIS information system are presented. Along with common techniques for the primary data source quality analysis, techniques for decomposition of expert data sources and pairwise comparison of ordered data sets are reviewed. We also discuss the use of empirical data for filtering large data collections by an acceptable difference between identical energy levels in primary data sources and in an empirical data source. Two types of the user interface for viewing the results of spectral data analysis and assessment of trust in expert data are considered. Original methods for the data source quality analysis and assessment of trust in expert data are briefly described. These methods are used for finding discrepancies between different spectral data types (primary, expert, and empirical).

Keywords: Molecular Spectroscopy, Spectral Resources, Quality Control of Spectral Data

1 Introduction

Assessment of the quality of information resources located on the Internet and representation of these resources in the form of software agents convenient for work are among currently important problems. Semantic Web (SW) technologies [1] were suggested for their solution. The tools for representing semantic annotations of resources turned out to be universal, while the techniques for assessment of the quality of resources, at least for scientific subject domains, are not universal.

Assessments of the quality of scientific resources can be divided into two groups. In the first group, the reliability of scientific resources connected with a mathematical model of a subject domain is assessed. For the validity check, criteria are used derived from the constraints imposed by the model on the entities described in the subject domain. In the second group, trust in resources is assessed, in particular, the influence of experts, presenting these resources.

Copyright © 2020 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR-WS.org/Vol-2790/paper24.pdf

When studying the states and transitions of molecules and atoms, large-volume numerical arrays are used; some of them contain parameters of several billion transitions for a molecule.

Virtual Atomic and Molecular Data Centre (VAMDC) [2-4] is currently the most famous spectral data center. It consolidates about 30 organizations and includes 39 databases, which include atomic (8 databases), molecular (22), and solid-state spectroscopy (2) data. The remaining DBs contain information about the kinetics of photochemical processes, collisions of molecules, atoms, and electrons, absorption cross sections during photo dissociation, and dissipative processes during electron collisions. Most of the atomic and molecular databases are components of information systems (IS) with web interfaces. All atomic molecular spectroscopy ISs provide researchers with lists of unique transitions, which can contain calculated transition parameters or mixtures of calculated and measured parameters. The quality of data in these systems is checked and assessed by the authors by way of comparison of their own results with those of other researchers. However, the results of these comparisons are described very briefly and selectively. The lack of explicit representation of the data quality analysis results and incomplete information about the analysis make it necessary to re-assess the data quality. This reassessment is time consuming and labor-intensive. Therefore, researchers have to decide based on the authoritative trust criterion in most cases. The disadvantage of the VAMDC site is the inaccessibility of the information resource quality analysis results.

Information system W@DIS [5] is a part of VAMDC; it includes primary, expert, and empirical spectral data and provides a complete description of the results of their quality assessment. These data describe more than hundred molecules (including isotopologues). A feature of W@DIS IS is the presence of complete results of the physical parameters and data sources quality analysis for all spectral data contained in it. All results of quality assessment are presented in the form of ontologies [6–9].

This paper continues the cycle of our works [10-12] on the study of various aspects of the analysis of the quality of spectral data. The second half of this paper partially overlaps with the extended abstract of the report [13]. In this work, we discuss the sequence of application of different methods for assessment of spectral information quality, features of these methods, and user interfaces for handling the results of spectral information quality analysis.

2 Data Model in Quantitative Spectroscopy. Data Quality Control

Before consideration of a data model for quantitative spectroscopy, we should emphasize the fact that the model suggested does not describe all facts related to the subject domain under study. However, it includes most data used in applied subject domains. The data model describes three data types.

Data related to measurements or calculations published in a particular work and their brief description together are called the primary data source. Primary experimental and theoretical data are interrelated, because not all values of the properties of transitions

and states are measurable, for example, quantum numbers, the values of which can be determined only by solving computational problems. In addition to primary data, composite data are also used in spectroscopy. Two composite data types are important for applications: empirical and expert data. The former are calculated by using a multiset of measured data, and the latter are a set of calculated, measured, and reference data. Thus, the data layer contains three data types: measurement data and their derivatives (empirical energy levels calculated on the basis of the Rydberg–Ritz principle [14]), calculation data, and expert data, which are a mixture of the first and second data types.

Figure 1 shows three groups of data and their relationship with experiment and calculations. Each data group is marked by a rounded rectangle; and the applications where these data are found, by rectangles. These applications are connected with six problems of quantitative spectroscopy [15] and the problem of expert data sets construction. The quality of the data included in three datasets (multiset of measured transitions and predicted and reference transitions) can only be assessed by formal criteria (validity), while expert datasets should be tested in accordance with trust criteria, since formal methods for constructing such datasets are to be developed. The computational method used for the formalization of data processing for constructing expert datasets in W@DIS was described in [16–18].

Since the sources of spectral information in the form of expert datasets are in great demand, expert data are also of interest in this work. There are three reasons for this. Firstly, the expert data on applied subject domains, the number of which exceeds several dozen, are contradictory. Note that data for a particular subject domain must meet certain requirements for their quality. Secondly, expert data are formed via informal manipulations, where the professional skills and preferences of experts play an important role. This is why different forms of publication criteria are a useful tool for the expert data quality assessment. Thirdly, the values derived from measurements and calculations are adapted when forming expert data used in solution of applied problems.

The analysis of the quality of C^0 — the multiset of transitions characteristics of which are measured and published for each molecule takes the central place in the data analysis. This set consists of several parts. The part C^0_A includes incorrectly identified transitions which are excluded from the quality analysis. The part C consists of correctly identified transitions, but can contain conflicting values of wavenumbers and other characteristics for identical transitions. For a number of molecules, Cantor sets of states (sets with all elements unique) are composed based on the multiset of transitions measured. The states in these sets are described by empirical energy levels E_R (for example, for a water molecule [18]), which can be used for the check of the corresponding set of transitions C. For this, all possible transitions (C_R), identical to the transitions from the set C, should be constructed from the empirical E_R levels. These transitions, which are identical to the transitions from C_R, form the set of transitions C^i_R. It consists of two subsets A^i and A_i, which include the transitions from C with the wavenumbers, differing from those of identical transitions from C_R by less or more than Δ cm^{-1}, respectively. Here Δ is the maximum permissible deviation between the wavenumbers of transitions from the set C^i_R and of identical transitions from the set C_R.

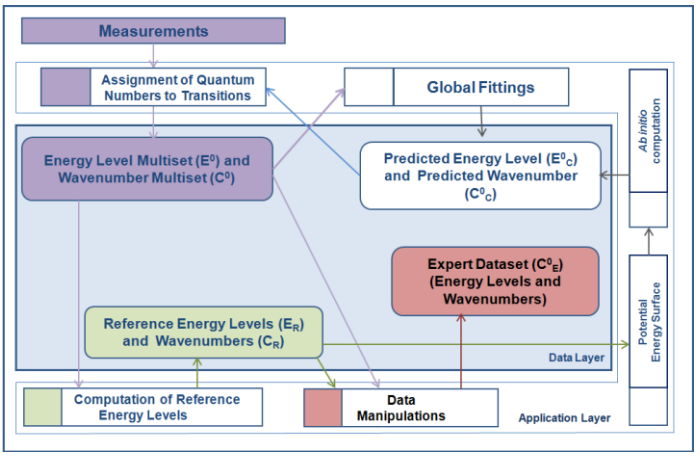


Figure 1. Schematic representation of the relationship between applications and associated input and output data [8].

Figure 2 shows the structure of the C^0 set. Let us write the obvious equalities: $C^i_R = A^i \cup A_i$, $C = C_A \cup C^i_R$, and $C^0 = C \cup C^0_A$ ($\cup$ is the union of sets).

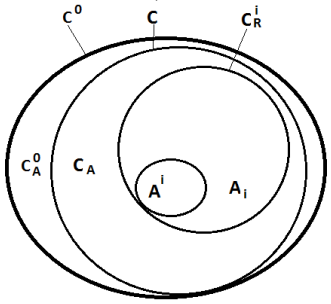


Figure 2. Structure of the multiset of transitions derived from the data quality analysis in the quantitative spectroscopy

The numbers of elements in the sets C^i_R and C_A do not change at a fixed number of the elements in the multiset C, but the number of elements in the sets A^i and A_i depends on Δ. This parameter is used as a quantitative measure of the permissible difference between the values of a wavenumber which describes a transition.

The quality analysis of data related to transitions from the set C_A is of greatest interest. If this set is Cantor, then a possibility of compiling a set of empirical energy levels from it is doubtful. Therefore, it is necessary to carry out measurements to expand the number of empirical energy levels in the set C^i_R.

The main information resources in W@DIS are sources of data on states and transitions. Each data source includes an uploaded numerical array or plot published and the properties of this array (plot).

Let D^0_n and D^0_{Em} denote the primary array of measured values of physical parameters and the expert data array, respectively. Figure 3 shows the order of the quality analysis

of an array uploaded (D^0_n or D^0_{Em}). Note that the above introduced multiset of transitions is the union of arrays:

$$C^0 = U^N_{n=1} D^0_{n.} \quad (1)$$

Hence, the equalities

$$C = U^N_{n=1} D_n,\ C^0_A = U^M_{m=1} D_m, \text{ etc.} \quad (2)$$

are true.

The ellipses in Fig. 3 show the parts of the dataset related to the parts of the multiset. The boxes represent applications which perform calculations during the data quality analysis. The in (out) arrows to a rectangle are associated with the input (output) of these applications. Plus and minus signs indicate that the output satisfies (+) or does not satisfy (-) the constraints of an application.

The difference in the data quality analysis for expert data in comparison with the previous one is the appearance of the Decomposition application during the second stage. This application outputs two data sets: D_{EDn} which includes physical parameter values close to published, and D_{AEDn} which include values which significantly differ from published ones. It is possible to work with both data groups obtained from decomposition in W@DIS. C_C is the set of primary computed transitions that is used during decomposition.

The data quality analysis results are part of the metadata for each data source in the W@DIS system. This metadata are represented in two forms: human readable and ontology for software agents.

3 General Description of the Collection of Spectral Line Parameters

The spectral data quality is analyzed in each paper on quantitative spectroscopy with the use of traditional methods. The division of spectral data into data sets, which relate to individual molecules and solutions to one of the seven spectroscopy problems in the W@DIS information system (IS) (http://wadis.saga.iao.ru) [15] allowed us to simplify the primary data analysis and to assess trust in expert data used in different application problems. Existing expert data created by different experts are contradictory. To resolve the contradictions, a decomposition method has been suggested to assess the trust in these data [10, 19]. The methods for the data quality analysis and assessment of trust in expert spectral data are briefly described in the report. Spectral data collections in W@DIS are associated with the output data of several dozens of atmospheric molecules [5]. The structure of a collection is identical to the structure of physical problems solved in molecular spectroscopy and includes data on energy levels, vacuum transitions, and spectral parameters of lines of individual molecules and their mixtures. Each part of a collection is assigned to the publication from which these data has been extracted.

Along with data sources, the system includes all publications related to data arrays accumulated in the collections.

Each data source is connected with a conventional set of metadata, which describes the set of properties of this source. Data on different molecules are not connected. This fact determines the branched modular structure of ontologies, each describing one of the modules. Every molecular collection is described by ontologies, which characterizes the information resources or physical quantities in this collection.

The main properties of data sources are those which describe the source quality analysis results. Data sources in the collections are typified. The following types are used for the classification: primary (measurements or calculations), expert, and empirical. The first three types are traditional for natural science data collections, while empirical data in different subject domains can be formed according to different principles. In spectroscopy, empirical data includes data related to energy levels and other physical quantities. Empirical energy levels are obtained from processing the complete set of energy levels derived from the inverse problem solution accounting the Rydberg–Ritz principle [9].

Authors of most publications use traditional methods for the analysis of spectral data quality: check of selection rules and calculation of standard deviations. When working with data collections, we use original methods for the data quality analysis and assessment of trust in expert data. These methods are briefly described below.

4 Methods of Spectral Data Quality Analysis

These methods can be divided into traditional and new [8, 10, 11]; we have introduced the latter for controlling the quality of primary and expert data. They are briefly described below.

4.1 Traditional Methods for Spectral Data Quality Analysis

During the analysis of spectral data quality, the states and transitions with incomplete set of quantum numbers are first distinguished. The second group of data excluded from the consideration includes transition duplicates in a data source.

The values of quantum numbers which characterize the energy levels and transitions between energy levels in a molecule are limited by the selection rules which follow from the mathematical model of a molecule. In W@DIS, forbidden transitions are marked and do not participate in any operations with data sources.

Paired relationships between data sources are estimated by the standard deviation or maximal deviation between the values of a physical quantity, identical states, or identical transitions compared. The acceptable deviation is determined by a researcher which uses the data to solve a problem.

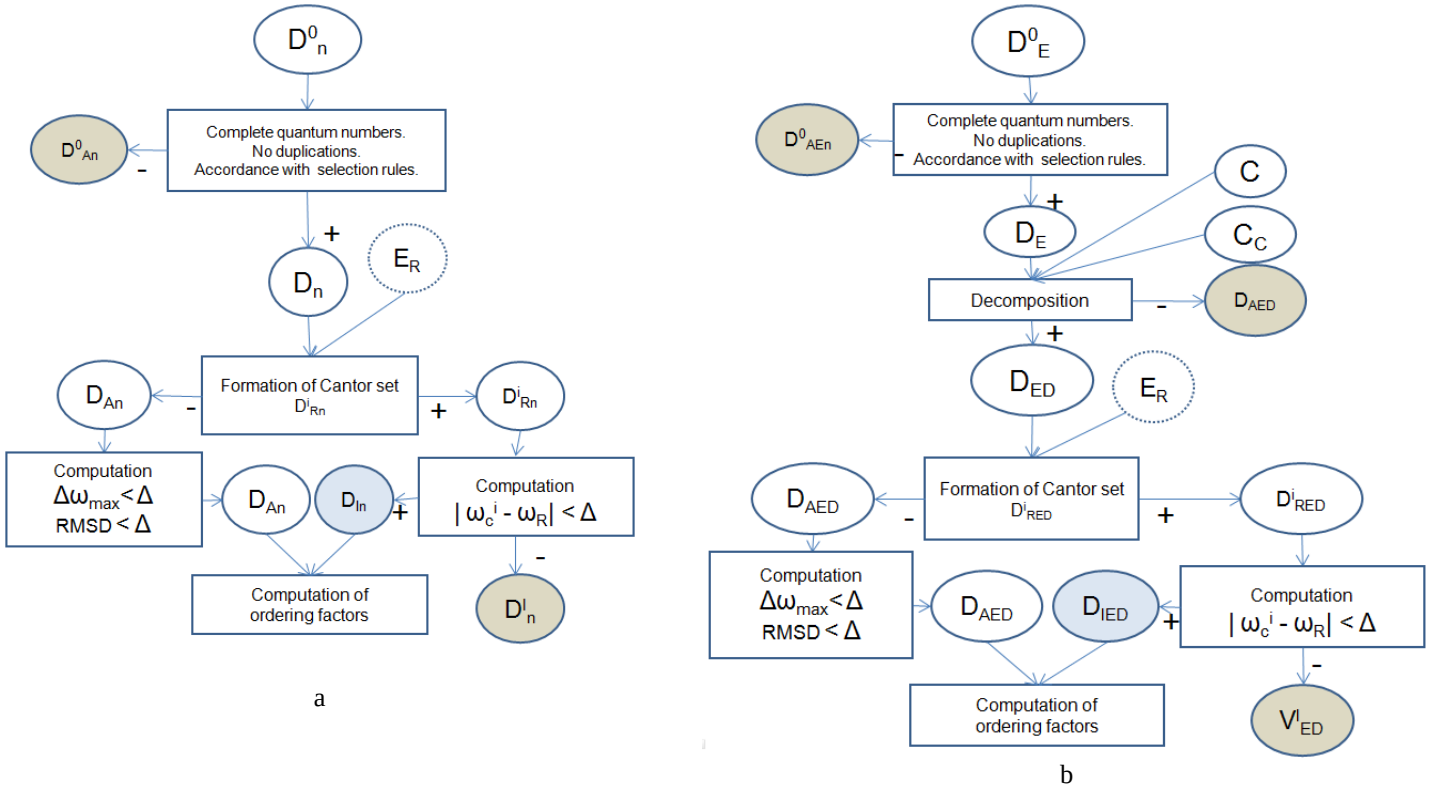


Figure 3. Sequence of actions in spectral data quality analysis: (a) measured data and (b) expert data

4.2 The Use of Empirical Data for the Collection Quality Control

Along with the above described methods, there is an approach which uses the empirical data calculated and allows improvement of the quality of spectral data in the collections. Such empirical data have been acquired for several dozens of molecules. This method is currently applied to data on all isotopologues of the water molecule in W@DIS. The allowable difference determined by an expert (0.035 cm^{-1}) between the energy levels of identical states taken from the data source, checked for quality, and from an empirical data source forms a filtering criterion. The results of application of this criterion are shown in Fig. 6.

The efficiency of this method depends on the number of states with empirical values of the energy levels E_R. The efficiency is maximal when the set C_A is empty.

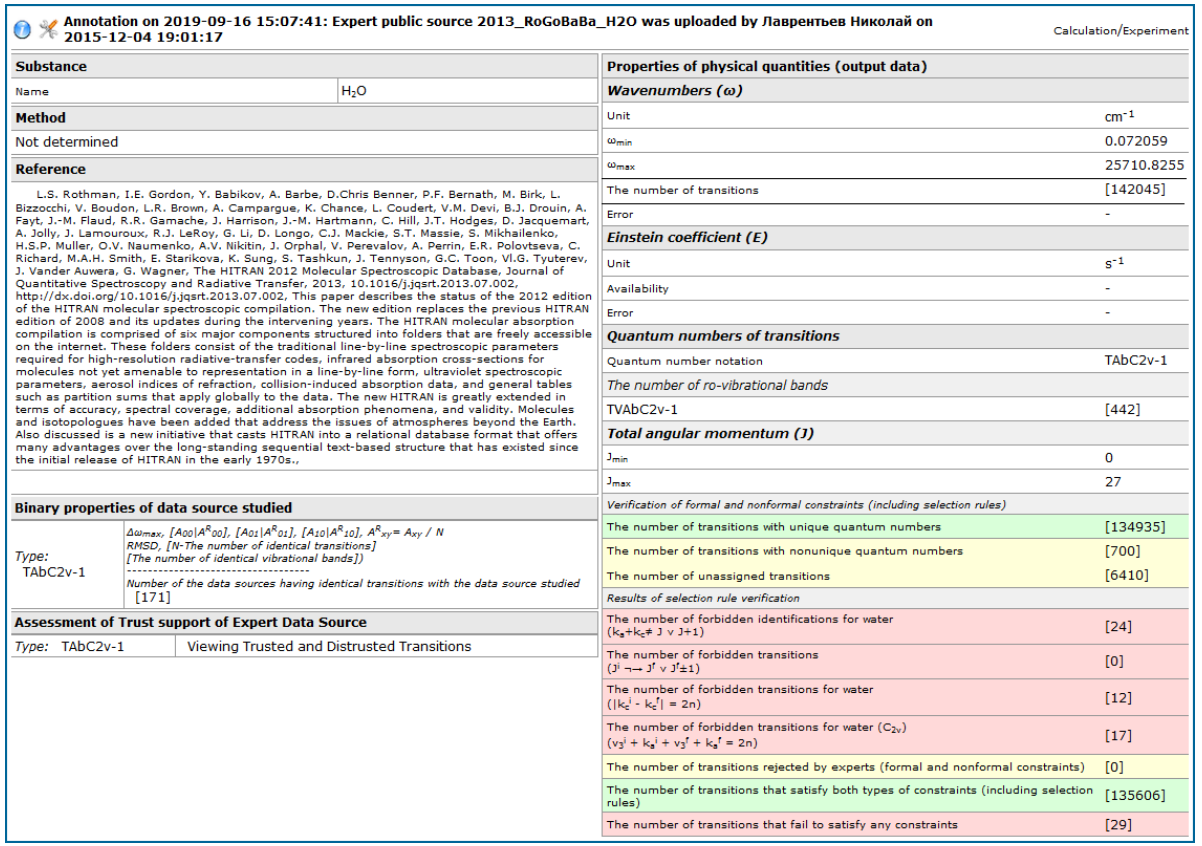
4.3 Pairwise Comparison of Ordered Spectral Data Arrays

The method [15] is used for identification of qualitative contradictions related to the sequence of states or transitions identified. It is based on two assumptions: the set of ideal states (transitions) is a Cantor set and it determines the ideal order of states (transitions); ideal values of physical quantities imply the ideal order of states (transitions). Three disorder factors [10] are calculated in the method, which allow one to estimate the different causes of arising contradictions. The software implementation of the method and its practical application faced the problem of long time required for the comparison of arrays of calculated data, including several billion transitions. The time required for the calculation of the factors exceeds several years.

Our recent studies have shown that the need in this method arises in the cases where traditional analysis techniques give a satisfactory quantitative result, but doubts remain about the degree of accuracy. In other words, this method should be applied in the cases where, for example, the spectral line centers should be determined with high accuracy in narrow spectral ranges.

4.4 Assessment of Trust in Expert Spectral Data

The above methods for the data quality analysis are applied to the primary data derived from measurements and calculations. Expert data are generated according to informal criteria, including subjective criteria, expediency, etc. Therefore, these data require another method for qualitative assessment. The method of expert data array decomposition we use is based on the following assumption: expert data should not differ significantly from the full set of primary data published. The decomposition by complete consistent (rigorous) or inconsistent set (non-rigorous method) is admissible. In spectroscopy, the non-rigorous method is required during the initial stage of study of spectral properties of molecules when there is no sufficiently good model or accurate measurement data. A researcher must choose numerical values based of the requirements for the task he solves.



Annotation on 2019-09-16 15:07:41: Expert public source 2013_RoGoBaBa_H2O was uploaded by Лаврентьев Николай on 2015-12-04 19:01:17

Calculation/Experiment

Substance

Name	H_2O

Method

Not determined

Reference

L.S. Rothman, I.E. Gordon, Y. Babikov, A. Barbe, D.Chris Benner, P.F. Bernath, M. Birk, L. Bizzocchi, V. Boudon, L.R. Brown, A. Campargue, K. Chance, L. Coudert, V.M. Devi, B.J. Drouin, A. Fayt, J.-M. Flaud, R.R. Gamache, J. Harrison, J.-M. Hartmann, C. Hill, J.T. Hodges, D. Jacquemart, A. Jolly, J. Lamouroux, R.J. LeRoy, G. Li, D. Longo, C.J. Mackie, S.T. Massie, S. Mikhailenko, H.S.P. Muller, O.V. Naumenko, A.V. Nikitin, J. Orphal, V. Perevalov, A. Perrin, E.R. Polovtseva, C. Richard, M.A.H. Smith, E. Starikova, K. Sung, S. Tashkun, J. Tennyson, G.C. Toon, Vl.G. Tyuterev, J. Vander Auwera, G. Wagner, The HITRAN 2012 Molecular Spectroscopic Database, Journal of Quantitative Spectroscopy and Radiative Transfer, 2013, 10.1016/j.jqsrt.2013.07.002, http://dx.doi.org/10.1016/j.jqsrt.2013.07.002, This paper describes the status of the 2012 edition of the HITRAN molecular spectroscopic compilation. The new edition replaces the previous HITRAN edition of 2008 and its updates during the intervening years. The HITRAN molecular absorption compilation is comprised of six major components structured into folders that are freely accessible on the internet. These folders consist of the traditional line-by-line spectroscopic parameters required for high-resolution radiative-transfer codes, infrared absorption cross-sections for molecules not yet amenable to representation in a line-by-line form, ultraviolet spectroscopic parameters, aerosol indices of refraction, collision-induced absorption data, and general tables such as partition sums that apply globally to the data. The new HITRAN is greatly extended in terms of accuracy, spectral coverage, additional absorption phenomena, and validity. Molecules and isotopologues have been added that address the issues of atmospheres beyond the Earth. Also discussed is a new initiative that casts HITRAN into a relational database format that offers many advantages over the long-standing sequential text-based structure that has existed since the initial release of HITRAN in the early 1970s.,

Binary properties of data source studied

Type: TAbC2v-1	*$\Delta\omega_{max}$, $[A_{00}\|A^R_{00}]$, $[A_{01}\|A^R_{01}]$, $[A_{10}\|A^R_{10}]$, $A^R_{xy}= A_{xy} / N$ RMSD, [N-The number of identical transitions] [The number of identical vibrational bands])* ---------- *Number of the data sources having identical transitions with the data source studied* [171]

Assessment of Trust support of Expert Data Source

Type: TAbC2v-1	Viewing Trusted and Distrusted Transitions

Properties of physical quantities (output data)

Wavenumbers (ω)	
Unit	cm^{-1}
ω_{min}	0.072059
ω_{max}	25710.8255
The number of transitions	[142045]
Error	-
Einstein coefficient (E)	
Unit	s^{-1}
Availability	-
Error	-
Quantum numbers of transitions	
Quantum number notation	TAbC2v-1
The number of ro-vibrational bands	
TVAbC2v-1	[442]
Total angular momentum (J)	
J_{min}	0
J_{max}	27
Verification of formal and nonformal constraints (including selection rules)	
The number of transitions with unique quantum numbers	[134935]
The number of transitions with nonunique quantum numbers	[700]
The number of unassigned transitions	[6410]
Results of selection rule verification	
The number of forbidden identifications for water ($k_a+k_c\neq J \vee J+1$)	[24]
The number of forbidden transitions ($J' \nrightarrow J'' \vee J''\pm1$)	[0]
The number of forbidden transitions for water ($\|k_c' - k_c''\| = 2n$)	[12]
The number of forbidden transitions for water (C_{2v}) ($v_3' + k_a' + v_3'' + k_a'' = 2n$)	[17]
The number of transitions rejected by experts (formal and nonformal constraints)	[0]
The number of transitions that satisfy both types of constraints (including selection rules)	[135606]
The number of transitions that fail to satisfy any constraints	[29]

Figure 4. Interface for viewing metadata of a data source. The left part contains the analysis results for binary relationships between data sources; the right bottom part contains the analysis results for selection rule checking.

Verification of formal and nonformal constraints (including selection rules)	
The number of transitions with unique quantum numbers	[134935]
The number of transitions with nonunique quantum numbers	[700]
The number of unassigned transitions	[6410]
Results of selection rule verification	
The number of forbidden identifications for water (k_a+k_c≠ J ∨ J+1)	[24]
The number of forbidden transitions (J^i ¬→ J^f ∨ J^f±1)	[0]
The number of forbidden transitions for water ($\|k_c^i - k_c^f\|$ = 2n)	[12]
The number of forbidden transitions for water (C_{2v}) ($v_3^i + k_a^i + v_3^f + k_a^f$ = 2n)	[17]
The number of transitions rejected by experts (formal and nonformal constraints)	[0]
The number of transitions that satisfy both types of constraints (including selection rules)	[135606]
The number of transitions that fail to satisfy any constraints	[29]

a

Binary properties of data source studied	
Type: TAbC2v-1	*$\Delta\omega_{max}$, [A_{00}|A^R_{00}], [A_{01}|A^R_{01}], [A_{10}|A^R_{10}], A^R_{xy}= A_{xy} / N* *RMSD, [N-The number of identical transitions]* *[The number of identical vibrational bands])* ---------- *Number of the data sources having identical transitions with the data source studied* [171]
Data Source	*Vacuum Wavenumbers*
2007_ZoOvShPo_H2O	*1.927e+3 [5229410|5.3e+1] [94676|9.7e-1] [97573|1.0e+0]* *7.937e+1 [97843]* *[428]*
2005_RoJaBaBe_H2O	*5.623e+2 [77851|2.4e+0] [7084|2.2e-1] [7535|2.3e-1]* *1.351e+1 [32070]* *[146]*
1995_PaHo_H2O	*8.238e-3 [0] [0] [0]* *6.140e-4 [228]* *[1]*

b

Figure 5. Fragments of the interface for viewing the analysis results for paired relationships (one–to–many) between data sources.

5 Interfaces for Viewing Data Quality Analysis Results

Two sets of interfaces are created in W@DIS for viewing the data quality analysis results. The first interface is intended for viewing metadata of a data source selected and its correlations with all other sources. The second interface is designed for description of paired relationships for the entire collection related to a molecule.

5.1 Comparison of a Selected Data Source with all the Sources which Include Identical States or Transitions (One-to-Many)

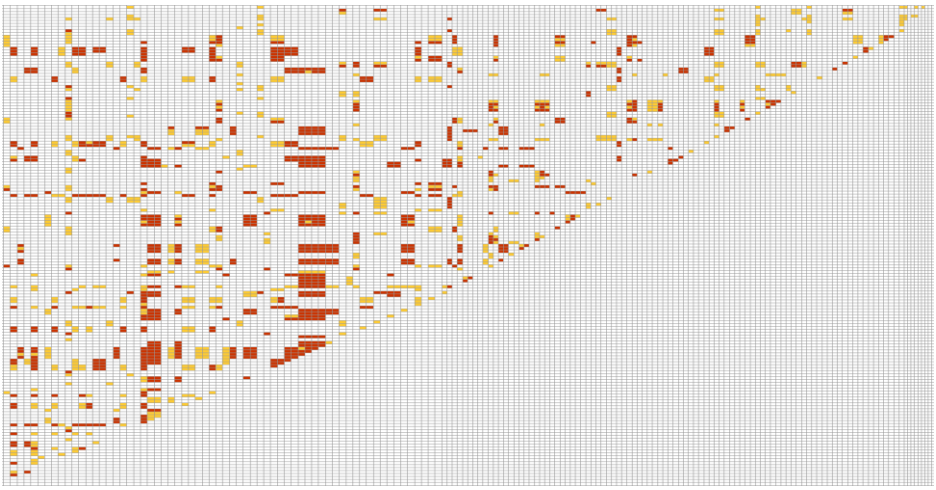
Figure 4 shows the interface for viewing metadata of expert data source [19]. It consists of four parts. A reference and a molecule are described in the upper part on the left, and physical quantities are described on the right. The bottom part of the interface shows the interfaces for viewing correlations between data sources and assessment of trust in the expert data array for the $H_2{}^{16}O$ molecule (on the left). A more detailed description is given in Fig. 5.

This comparison is carried out when describing a data source selected, but it can also be performed as refinement of the comparison (many–to–many) when describing a collection. Figure 4a shows that there are no forbidden transitions according to the selection rule checking; 6410 transitions were not identified. Fig. 5b shows a part of the list (comparison was made with 171 data sources), where the results disagree with the calculation data [20]. The discrepancy with the expert data array [21] is much smaller; the quantitative agreement with the following four works is good, while the ordering factors are nonzero and one line in the spectrum is disordered [22].

5.2 Complete Pairwise Comparison of Data Sources that Have Identical States or Transitions (Many–to-Many)

The results of comparison of all data sources can be presented in one viewing interface with quantitative characteristics or without them, i.e. qualitatively. Such a qualitative representation is shown in Fig. 6a, where pairs of sources are shown in red if the criterion chosen by a researcher is violated. This representation allows one to visually assess the number of good or bad pairs. Fig. 6b shows a fragment of the interface, where the quantitative coincidence of the data compared is obvious.

Figure 6a shows a part of a symmetric matrix with the qualitative results of estimation of the maximal difference between the frequencies in two published arrays of transitions, which are to be compared and include parameters of identical transitions, in the cells. If the maximal difference is less than 0.035 cm^{-1}, then the corresponding cell is light gray; if the cell is white, this means that the compared arrays have no identical transitions, and if the cell color is dark gray, then the maximal difference is higher than 0.035 cm^{-1}. Note that the number of cells with the high difference is quite large (slightly less than half of all data). The arrays of values published have passed the first stage of data filtering.



a

Источники данных				#36	#35	#34	#33	#32	#31	#30	#29	#28	#27
1971_HuDa_CH4	1	9	#1			4.89e-1, 2			3.18e+0, 79				
1971_HuPo_CH4	0	7	#2	4.61e+1, 9									
1972_BaSuHuPl_CH4	0	9	#3			1.18e+0, 4			7.00e+2, 111				
1972_Bobin_Ch4			#4										
1972_Botineau_CH4	1	4	#5										
1972_HuPoVaAm_CH4	0	1	#6										
1973_HoGeOz_CH4	2	0	#7										
1973_Susskind_CH4	0	5	#8	7.59e+1, 52									
1974_TaDaPo_CH4	0	7	#9						4.75e+0, 13				
1975_HoGeOz_CH4	2	0	#10										
1975_Pine_CH4_B	0	6	#11			1.18e+0, 3			3.73e-1, 78				
1975_Pine_CH4_BF	1	4	#12			4.87e-1, 1			8.08e-1, 26				
1975_TaDaPoGu_CH4	0	2	#13						1.17e+1, 18				
1976_OzLeGe_CH4	1	0	#14										
1977_ToBrHu_CH4	0	3	#15						1.75e+0, 15				
1979_DaPiRo_CH4	0	2	#16						8.06e-1, 6				
1979_GrRoPi_CH4	1	1	#17			1.15e+0, 2			1.37e-2, 129				

b

Figure 6. Interface for viewing the analysis results for paired relationships (many–to–many) between data sources (comparison of sources from the methane molecule data collection): (a) qualitative presentation (pairs which do not satisfy the criterion are shown in dark gray (red)); (b) fragment of the interface with matrix quantitative representation (pairs which satisfy the criterion are shown in light gray (yellow)); the first number is the maximal difference, and the second is the number of identical transitions in a pair.

5.3 The Use of Empirical Data to Control the Quality of the Collection of Spectral Data on the Water Molecule

Figure 7 shows the quality analysis results for the multiset of H_2O transitions measured, which have been tested by the criterion of maximal deviation between the values of wavenumbers of a pair of identical transitions $\Delta < 0.035$ cm^{-1}.

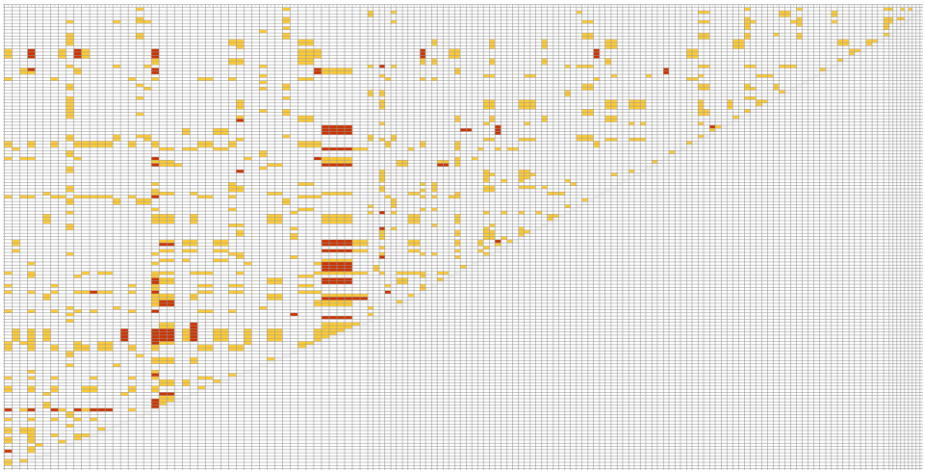


Figure 7. The analysis results for paired relationships (many-to-many) between data sources (qualitative comparison of the sources of data collection on the water molecule)

Table 1. Number of transitions in different parts of the multiset C^0.

Molecule $\mathbf{H_2{}^{16}O}$	The number of measured transitions	
Transition groups and associated arrays	Total transitions	of which are unique
C^0 - The number of transitions uploaded to the collection	312366	-
$C^0{}_A$ - Transitions with incorrect quantum numbers	19932	-
C - Transitions with correct quantum numbers	292434	103324
C_A - Transitions in which at least one state (upper or lower) is in E_R	1400	837
C^i_R - Transitions in which both states (upper and lower) are in E_R	291034	102487
A_i - Identical transitions for which $\lvert\omega_{CiR} - \omega_{CR}\rvert < \Delta$	281561	100270
A^i - Identical transitions for which $\lvert\omega_{CiR} - \omega_{CR}\rvert > \Delta$	9473	5302

Figure 7 shows the results of comparison of the same data arrays, but after the second stage of filtering, during which transitions with the wavenumbers from the C array

which differed from the corresponding wavenumbers of the identical transitions in the C^i_R array were rejected. The number of dark gray cells in Fig. 7 is much less than in Fig. 6a.

Table 1 presents number of transitions in different parts of the multiset C^0. They are calculated for the transitions of the main isotopologue of the water molecule. In the table one can check the following equalities from Figure 2: the number of transitions in C is equal to the sum of transitions in the sets C_A and C^i_R and the number of unique transitions in C is equal to the sum of unique transitions in the sets C_A and C^i_R.

Details of the study on improvement of the quality of data on the water molecule are given in [23]. Here, we present only the figure which characterizes the collection data quality before and after the filtration.

5.4 Interface for Viewing the Results of Assessment of Trust in Expert Data

Fig. 8 shows a fragment of the interface which represents a part of the results of trust assessment. This part characterizes the decomposition results for a complete set of computed and measured data.

Assessment of Trust support of Expert Data Source										
Type: TAbC2v-1	Viewing Trusted and Distrusted Transitions									
Decomposition Group. *(Note that RF, L.W. and S.W. IR mean radio-frequency, long-wave and short-wave region, respectively.)*	*Vacuum Wavenumbers*									
	RF	*Microwave*	*Far IR*	*L.W. IR*	*Middle IR*	*S.W. IR*	*Near IR*	*Visible*	*Near UV*	*All Regions*
Assessment of Trust Based on Primary Computed Data										
Distrusted Vacuum Wavenumbers	*0.072* *0.072* *[1]*	*0.117* *9.921* *[119]*	*10.476* *629.922* *[3822]*	*630.028* *1249.709* *[2046]*	*1251.059* *3299.796* *[10266]*	*3300.008* *6999.976* *[18466]*	*7000.013* *13499.930* *[40450]*	*13500.043* *23999.560* *[22130]*	*24001.250* *25336.555* *[194]*	*0.072* *25336.555* *[97494]*
Assessment of Trust Based on Primary Measured Data										
Distrusted Vacuum Wavenumbers	*0.072* *0.072* *[1]*	*0.117* *9.921* *[101]*	*10.098* *625.825* *[961]*	*649.426* *1025.966* *[5]*	*1526.119* *3251.883* *[21]*	*3393.330* *6997.888* *[508]*	*7004.259* *13499.872* *[6734]*	*13500.234* *23999.000* *[17701]*	*24000.379* *25689.403* *[211]*	*0.072* *25689.403* *[26243]*
Assessment of Trust Based on Primary Measured and Computed Data										
Distrusted Vacuum Wavenumbers	*0.072* *0.072* *[1]*	*0.117* *9.921* *[101]*	*10.476* *625.825* *[797]*	*649.426* *1025.966* *[4]*	*1526.119* *3184.313* *[20]*	*3393.330* *6997.888* *[451]*	*7004.290* *13499.872* *[5377]*	*13500.234* *23923.215* *[4128]*	*24831.528* *25229.911* *[17]*	*0.072* *25229.911* *[10896]*

Figure 8. A fragment of the interface for viewing the results of the assessment of trust in the expert HITRAN data array [20].

The expert source contains 135606 transitions in total. The lines “Distrusted Vacuum Wavenumbers” characterize the number of transitions with low trust in different spectral ranges and over the entire spectral region. These transitions make 8% of the total number of transitions. Most low-trust transitions (87%) are in the near-IR and visible regions.

Conclusions

We describe a collection of spectral data in W@DIS information system and the methods used in it for the analysis of spectral data quality. A data model in quantitative spectroscopy and a group of applications for data acquisition are presented. Three data groups are distinguished: data derived from measurements, from calculations, and their combination in the form of expert spectral data. Different parts of the multiset of transitions, which consists of numerical arrays extracted from publications, are described. The sequence of actions for the data quality analysis is explained. The quality assessment techniques commonly used in spectroscopy are briefly described, as well as three methods for the spectral data quality analysis developed by the authors for large arrays of conflicting spectral data. For better visual representation of the results, graphical user interfaces for working with the results of data quality assessment are exemplified. The percentage ratio of the parts of the multiset is shown, with the multiset of transitions of the main water molecule isotopologue as an example. In particular, we have shown that less than 1.5% of the data remains after data filtering in the automatic spectral data quality analysis, which require additional processing by experts.

Acknowledgement. The main part of the work was performed within project № AAAA-A17-117021310147-0.

References

1. Berners-Lee, T., Hendler, J., Lassila,O.: The Semantic Web. Scientific American (2001)
2. Dubernet, M.L., Boudon, V., Culhane, L., Dimitrijevic, M., Fazliev, A., Joblin, C., Kupka, F., Leto, G., Le Sidaner, P., Loboda, P., Mason, H., Mason, N., Mendoza, C., Mulas, G., Millar, T., Nunez, L., Perevalov, V., Piskunov, N., Ralchencko, Y., Rothman, L., Ryabchikova, T., Ryabtsev, A., Sahal-Brechot, S., Schmitt, B., Schlemmer, S., Tennyson, J., Tyuterev, V., Walton, N., Wakelam, V., Weiss, W., Zeippen, C.: Virtual Atomic and Molecular Data Centre. Journal of Quantitative Spectroscopy and Radiative Transfer, v.111, no.15, 2151-2159 (2010)
3. Dubernet-Tuckey, M.-L., Antony, B., Ba, Y.-A., Babikov, Yu., Bartschat, K., Boudon, V., Braams, B., Chung, H.-K., Daniel, F., Delahaye, F., Del Zanna, G., de Urquijo, J., Dimitrijevic, M., Domaracka, A., Doronin, M., Drouin, B., Endres, C., Fazliev, A., Gagarin, S., Gordon, I., Gratier, P., Heiter, U., Hill, C., Jevremovic, D., Joblin, C., Karsprzak, A., Krishnakumar, E., Leto, G., Loboda, P.A., Louge, T., Maclot, S., Marinkovic, B., Kemper, A.M., Marquart, T., Mason, H., Mason, N., Mendoza, C., Mihajlov, A., Millar, T., Moreau, N., Mulas, G., Pakhomov, Yu., Palmeri, P., Pancheshnyi, S., Perevalov, V.I., Piskunov, N., Postler, J., Quinet, P., Sánchez, E.L.Q., Ralchenko, Yu., Rhee, Yong-Joo, Rixon, G., Rothman, L., Roueff, E., Ryabchikova, T., Sahal-Brechot, S., Scheier, P., Schlemmer, S., Schmitt, B., Stempels, E., Tashkun, S., Tennyson, J., Tyuterev, V., Vujcic, V., Wakelam, V., Walton, N., Zatsarinny, O., Zeippen, C., Zwölf, C.M.: The Virtual Atomic and Molecular Data Centre (VAMDC) Consortium. Journal of Physics B: Atomic, Molecular and Optical Physics 49(7):074003 (2016)
4. Albert, D., Antony, B.K., Ba, Yaye Awa, Babikov, Yu.L., Bollard, P., Boudon, V., Delahaye, F., Del Zanna, G., Dimitrijevíc, 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., Jevremovíc, D., Joblin, C., Kwon, D.-H. Kochanov, R.V., Krishnakumar, E., Leto, G., Loboda, P.A., Lukashevskaya, A.A., Lyulin, O.M., Marinkovíc, 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, Yu., 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., Sreckovíc, V.A., Stempels, E., Tashkun, S.A., Tennyson, J., Tyuterev, V.G., Vastel, C., Vujcíc, V., Wakelam, V. Walton,, N.A., Zeippen C. and Zwöl C.M.: A Decade with VAMDC: Results and Ambitions, Atoms, 8(4), 76 (2020) https://doi.org/10.3390/atoms8040076

5. Akhlyostin, A., Apanovich, Z., Fazliev, A., Kozodoev, A., Lavrentiev, N., Privezentsev, A., Rodimova, O., Voronina, S., Csaszar, A.G., Tennyson, J.: The current status of the W@DIS information system. In: Matvienko, G., Romanovskii, O. (eds.) Proc. SPIE of 22-nd International Symposium Atmospheric and Ocean Optics: Atmospheric Physics, 10035, 100350D (2016)
6. Privezentsev A.I., Tsarkov D.V., Fazliev A.Z.: Computed knowledge base for description of information resources of molecular spectroscopy. 3. Basic and applied ontologies. Digital Library Journal 15(2) (2012)
7. Voronina, S.S., Privezentsev, A.I., Tsarkov, D.V., Fazliev, A.Z.: An Ontological Description of States and Transitions in Quantitative Spectroscopy. In: Proc. of SPIE XX-th Inter. Symp. Atmosp. Ocean Optics: Atmospheric Physics, v. 9292, 92920C (2014)
8. Fazliev, A., Privezentsev, A., Tsarkov, D., Tennyson, J.: Ontology-Based Content Trust Support of Expert Information Resources in Quantitative Spectroscopy. In: Knowledge Engineering and the Semantic Web, CCIS, v. 394, Springer Berlin Heidelberg, pp.15-28 (2013)
9. Watson, J.K.G.: The Use of Term-Value Fits in Testing Spectroscopic Assignments. Journal of Molecular Spectroscopy 165:283-290 (1994)
10. Akhlyostin, A., Lavrentiev, N., Privezentsev, A., Fazliev, A.: Computer knowledge bases for describing information resources in molecular spectroscopy. 5. Expert data quality. Rus. Dig. Libr. J. 16(4) (2013)
11. Lavrentiev, N., Makogon, M., Fazliev, A.: Comparison of the HITRAN and GEISA spectral databases taking into account the restriction on publication of spectral data. Atmos. Ocean. Opt. 24(5):436-451 (2011)
12. Fazliev A., Privezentsev A., Tsarkov D., Tennyson J.: Ontology-Based Content Trust Support of Expert Information Resources in Quantitative Spectroscopy. Knowledge Engineering and the Semantic Web, CCIS / Eds: P. Klinov, D. Mouromtsev. Springer Berlin, Heidelberg. 394:15-28 (2013)
13. Akhlestin A.Yu., Lavrentiev N.A., Kozodoev A.V., Fazliev A.Z., Privezentsev A.I., Kozodoeva E.M.: Data Quuality Assesssment in Large Collections of Spectral Data, In: Computing resources. Digital twins and big data. (DICR-2019). Proceedings of the XVII International Conference. P. 214-222. (2019) DOI: 10.25743/ICT.2019.25.66.032
14. Bykov, A.D., Voronin, B.A., Kozodoev, A.V., Lavrent'ev, N.A., Rodimova, O.B., Fazliev, A.Z.: Information system for molecular spectroscopy. 1. Structure of information resources. Atmosp. Oceanic Optics, 17:816-820 (2004)
15. Kozodoev, A.V., Fazliev, A.Z.: Information system for molecular spectroscopy. 2. Array operations for transformation of data on spectral line parameters. Atmos. Oceanic Optics, 18:680-684 (2005)

16. Kozodoev, A.V., Kozodoeva, E.M.: Extensible module «Unary operations» in the information system «Molecular spectroscopy». Vest. NSU. Inform. Technol., 13:46-54 (2015)
17. Kozodoev A.V., Kozodoeva E.M.: The binary operations in the information system «Molecular Spectroscopy». Vest. NSU. Inform. Technol., v. 16:70-77 (2018)
18. Tennyson, J., Bernath, P.F., Brown, L.R., Campargue, A., Császár, A.G., Daumont, L., Gamache, R.R., Hodges, J.T., Naumenko, O.V., Polyansky, O.L., Rothman, L.S., Vandaele, A. C., Zobov, N.F., Al Derzi, A.R., Fabrie, C., Fazliev, A., Furtenbacher, T., Gordon, I.E., Lodi, L., Mizus, I.: IUPAC Critical Evaluation of the Rotational-Vibrational Spectra of Water Vapor. Part III. Energy Levels and Transition Wavenumbers for $H_2{}^{16}O$, Journal of Quantitative Spectroscopy and Radiative Transfer, 117:29-58 (2013)
19. Rothman, L., Gordon, I., Babikov, Y., Barbe, A., Chris Benner, D., Bernath, P., Birk, M., Bizzocchi, L., Boudon, V., Brown, L., Campargue, A., Chance, K., Coudert, L., Devi, V., Drouin, B., Fayt, A., Flaud, J.-M., Gamache, R., Harrison, J., Hartmann, J.-M., Hill, C., Hodges, J., Jacquemart, D., Jolly, A., Lamouroux, J., LeRoy, R., Li, G., Longo, D., Mackie, C., Massie, S., Mikhailenko, S., Muller, H., Naumenko, O., Nikitin, A., Orphal, J., Perevalov, V., Perrin, A., Polovtseva, E., Richard, C., Smith, M., Starikova, E., Sung, K., Tashkun, S., Tennyson, J., Toon, G., Tyuterev, V., Vander Auwera, J., Wagner, G.: The HITRAN 2012 Molecular Spectroscopic Database. Journal of Quantitative Spectroscopy and Radiative Transfer, 130:4-50 (2013)
20. Zobov, N., Ovsyannikov, R., Shirin, S., Polyansky, O.: The assignment of quantum umbers in the theoretical spectra of the $H_2{}^{16}O$, $H_2{}^{17}O$, and $H_2{}^{18}O$ molecules calculated by variational methods in the region 0–26 000 cm^{-1}. Optics and Spectroscopy, v. 102(3) (2007). DOI: 10.1134/S0030400X07030058
21. Rothman, L., Jacquemart, J., Barbe, A., Benner, D., Birk, M., Brown, L., Carleer, M., Chackerian Jr., C., Chance, K., Coudert, L., Dana, V., Devi, V., Flaud, J.-M., Gamache, R., Goldman, A., Hartmann, J.-M., Jucks, K., Maki, A., Mandin, J.-Y., Massie, S., Orphal, J., Perrin, A., Rinsland, C., Smith, M., Tennyson, J., Tolchenov, R., Toth, R., Vander Auwera, J., Varanasi, P., Wagner, G.: The HITRAN 2004 molecular spectroscopic database. J. Quant. Spect. Rad. Transfer, v. 96:139–204 (2005)
22. Paso, R., Horneman, V.-M.: High-resolution rotational absorption spectra of $H_2{}^{16}O$, $HD^{16}O$, and $D_2{}^{16}O$ between 110 and 500 cm^{-1}. Journal of Optical Society of America, B, 12:1813-1838 (1995), DOI:10.1364/JOSAB.12.001813.
23. Kozodoev, A., Kozodoeva, E.: Assessment of reliability of spectral data on H_2O, H_2S, SO_2, C_2H_2, and NH_3 molecules from empirical sets of energy levels. Abstracts of XXV Intern. Symp. on Atmospheric and Ocean Optics. Atmospheric Physics (2019).