Systematization of Tabular and Graphical Resources in Quantitative Spectroscopy

© N.A. Lavrentiev © A.I. Privezentsev © A.Z. Fazliev
Institute of Atmospheric Optics SB RAS,
Tomsk, Russia
lnick@iao.ru remake@iao.ru faz@iao.ru

Abstract. An approach to the formation of applied ontologies in data intensive subject domains with predominant tabular and graphical forms of data representation is suggested. Sources of data and of information about data in tabular and graphical forms are described. Using the quantitative spectroscopy as an example, an approach is presented to the formation of semantic annotations characterizing these sources. The main types of sources and methods for controlling the spectral data quality are described. Using scientific graphics in the spectroscopy of molecular complexes as an example, an approach to the solution of the problem of reduction and classification of graphical resources for searching for elementary plots in the subject domain is described. The role of ontology metrics in the comparison between data collections is discussed.

Keywords: big data systematization, quantitative spectroscopy, applied ontologies.

1 Introduction

Research results in tabular and graphical forms take a significant part in publications related to data intensive subject domains. Usually, when processing such publications by search agents, this part of information resources is ignored. The reason is due to the lack of universal software, which allows describe of such resources from different subject domains.

The implementation of search for information about tabular and graphical resources was started in the 1990s using metadata integrated into html-pages. The creation of Semantic Web technology was declared in the early 2000s [1]; its aim was replacing traditional metadata by semantic annotations. No total transition to semantic annotations occurred, since, on the one hand, the introduction of new technologies turned out to be a complicated process and, on the other hand, there was no demand for detailed queries that gave near-unambiguous answers.

During the initial stage of the creation of the Web technologies, the volume of unscientific resources significantly exceeded the amount of scientific resources. Since the end of the 2000s, the situation has begun to change and the volume of scientific data has begun to grow catastrophically. In future, these data exceed all other resources [2]. Scientific information resources are represented on the Internet in publications (files), data collections (databases), subject domain ontologies (knowledge bases), etc. Below we mainly focus on scientific papers and their systematization. This part of the resources is chosen, on the one hand, because of their traditional use in research, and on the other hand, because of a need in searching for scientific resources with a highly detailed query. Note that already in the middle of the 2000s, attempts were made in several subject domains to systematize non-textual parts of scientific resources [3-5]. Methods for systematization in our work are detailed on examples from quantitative spectroscopy.

We have systematized sets of spectral data on spectroscopy during the past 15 years. Semantic annotations of these data sets have become a part of applied ontologies characterizing one of the basic properties of these sets, that is, the trust in these data [6]. We digitized tables and plots representing the parameters of spectral lines and spectral functions. The digitization of the tables was needed for the control of expert spectral data quality, and the digitization of spectral functions was caused by the need to have spectral information in the cases where there were no high resolution results, and also for their usage for controlling the asymptotic behavior of the calculated data.

We constructed applied ontologies that characterize the quality of information resources on molecular spectroscopy [7], states and transitions of atmospheric molecules [8] and graphical resources on spectroscopy [9]. The ontologies created characterize tabular data that describe the spectral lines studied during the past 80 years. In the first thirty years of this period, publications, along with a small number of data tables, included many scientific plots describing spectral functions. Creation of Fourier spectrometers in the late 1960s initiated the appearance of many numerical arrays of precise data on spectral lines parameters, and graphical representation of spectral data was replaced by tabular representation in high-resolution quantitative spectroscopy in subsequent years.

Nevertheless, there are spectroscopy domains where it is difficult to achieve a high resolution of the spectral parameters with the help of modern experimental techniques. For example, the continuum absorption,

Proceedings of the XX International Conference "Data Analytics and Management in Data Intensive Domains" (DAMDID/RCDL'2018), Moscow, Russia, October 9-12, 2018

important in the study of planetary and exoplanetary atmospheres; spectral properties of weakly bonded molecular complexes and molecular spectral functions in the UV region necessary for quantitative description of photochemical reactions in the gaseous phase. In these subject domains, the amount of spectral information contained in scientific graphics significantly exceeds the amount of information represented in the tabular form.

In this work, we discuss the models and features of tabular and graphical representations of data in scientific publications, define the primary and composite data sources, information sources, elementary and composite plots and figures. In the final part of the article, we estimate the metrics of the created ontologies on quantitative spectroscopy.

2 Features of tabular and graphical representations of resources in publications

2.1 Publication model

Publications are the most common means for storage, communication, and analysis of the scientific information. Traditionally, scientific papers include text in a natural language, mathematical equations, chemical reactions, physical formulas, tables, plots, figures, etc. To find the information requested by a user, the text part is mainly used. In many subject domains, data arrays, which are solutions to computational problems, measurements or observations, are used in tabular and graphical representations. Every such solution is a part of a paper that contains a large number of typed facts. Equations, formulas, and sets of reactions are much more abstract resources, since most of them have no unique names and their annotation requires a certain level of professional training.

To form the part of semantic annotations that characterizes tabular and graphical resources of a paper in a simple case, one can take into account the description of properties of the domain problem solutions. Note that the current trend is creation of supplementary materials to papers, many of which contain additional data in the tabular and/or graphical forms.

The solution of a computational problem is a data array supplemented by a set of properties of this array; it can represent a more accurate formal model of one or another part of a paper. The specification of the set of properties is determined by the problems of searching for information resources, which are of interest to researchers of the given subject domain.

The choice of a publication model for collections of data arrays represented in tabular and graphical forms is caused by the task of automatic cataloging of such informational resources in a subject domain. Our collection of papers on the quantitative spectroscopy already exceeds 12,000 publications relating to the period from 1898 till the present. The model of the subject domain chosen by us [8] contains solutions of seven spectroscopic problems that are of decisive importance for such applied subject domains, as astronomy, atmospheric optics, spectroscopy, etc.

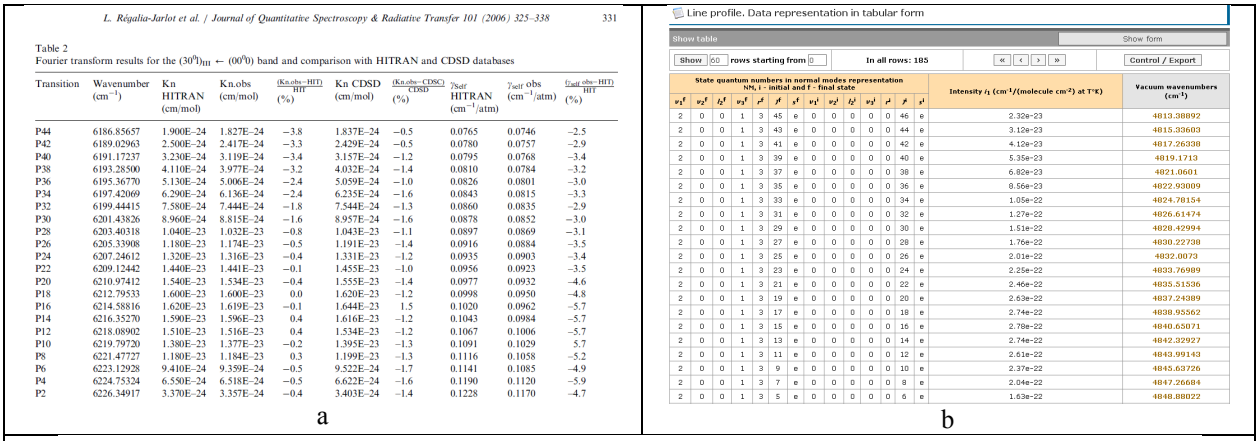
Tables in the publications contain not only data arrays, but also scientific graphics. Graphical resources in scientific subject domains can be divided into two parts: mathematical plots (usually 2- and 3D) and figures (raster graphics and graphics represented by data arrays). Today, digital images of scientific graphics appear in a number of journals in supplementary materials, which makes possible the quantitative comparison of graphics with less cost.

2.2 Tabular representation

Intensive use of numerical data led to a wide variety of forms of tabular representation. Tabular data in the paper text and in plain-text files contain data arrays with positional formatting with whitespace characters or formatting with separating symbols, so-called CSV files (Comma-Separated Values). The form of a table does not impose restrictions on metadata to the data arrays. The subject domain model chosen in a specific information system allows one to distinguish the structure of the intension of semantically significant data arrays in a tabular form. Thus, not all information published in the tabular form should be semantically annotated, but only the information necessary for W@DIS information system.

In journals, the tabular data representation is still used in spectroscopy, but the volume of spectral data there has decreased significantly; most of the information resources presented in the tabular form are concentrated in supplementary materials. Note that the number of plots in papers was much higher than of tables in the first half of the 20th century.

In the W@DIS information system described below, data arrays extracted from tables published in scientific papers are the main resources. Figure 1 shows some stages of the formation of these resources. Figure 1a shows a fragment of a table from a paper; Fig. 1b gives a typical representation of data from tables in W@DIS, and Fig. 1c shows metadata that are automatically generated when importing data published into the IS.



L. Régalia-Jarlot et al. / Journal of Quantitative Spectroscopy & Radiative Transfer 101 (2006) 325–338 331

Table 2
Fourier transform results for the $(30^01)_{III} \leftarrow (00^00)$ band and comparison with HITRAN and CDSD databases

Transition	Wavenumber (cm^{-1})	Kn HITRAN (cm/mol)	Kn.obs (cm/mol)	$\frac{(Kn.obs-HIT)}{HIT}$ (%)	Kn CDSD (cm/mol)	$\frac{(Kn.obs-CDSC)}{CDSD}$ (%)	γ_{Self} HITRAN (cm^{-1}/atm)	γ_{Self} obs (cm^{-1}/atm)	$\frac{(\gamma_{Self}\ obs-HIT)}{HIT}$ (%)
P44	6186.85657	1.900E−24	1.827E−24	−3.8	1.837E−24	−0.5	0.0765	0.0746	−2.5
P42	6189.02963	2.500E−24	2.417E−24	−3.3	2.429E−24	−0.5	0.0780	0.0757	−2.9
P40	6191.17237	3.230E−24	3.119E−24	−3.4	3.157E−24	−1.2	0.0795	0.0768	−3.4
P38	6193.28500	4.110E−24	3.977E−24	−3.2	4.032E−24	−1.4	0.0810	0.0784	−3.2
P36	6195.36770	5.130E−24	5.006E−24	−2.4	5.059E−24	−1.0	0.0826	0.0801	−3.0
P34	6197.42069	6.290E−24	6.136E−24	−2.4	6.235E−24	−1.6	0.0843	0.0815	−3.3
P32	6199.44415	7.580E−24	7.444E−24	−1.8	7.544E−24	−1.3	0.0860	0.0835	−2.9
P30	6201.43826	8.960E−24	8.815E−24	−1.6	8.957E−24	−1.6	0.0878	0.0852	−3.0
P28	6203.40318	1.040E−23	1.032E−23	−0.8	1.043E−23	−1.1	0.0897	0.0869	−3.1
P26	6205.33908	1.180E−23	1.174E−23	−0.5	1.191E−23	−1.4	0.0916	0.0884	−3.5
P24	6207.24612	1.320E−23	1.316E−23	−0.4	1.331E−23	−1.2	0.0935	0.0903	−3.4
P22	6209.12442	1.440E−23	1.441E−23	−0.1	1.455E−23	−1.0	0.0956	0.0923	−3.5
P20	6210.97412	1.540E−23	1.534E−23	−0.4	1.555E−23	−1.4	0.0977	0.0932	−4.6
P18	6212.79533	1.600E−23	1.600E−23	0.0	1.620E−23	−1.2	0.0998	0.0950	−4.8
P16	6214.58816	1.620E−23	1.619E−23	−0.1	1.644E−23	1.5	0.1020	0.0962	−5.7
P14	6216.35270	1.590E−23	1.596E−23	0.4	1.616E−23	−1.2	0.1043	0.0984	−5.7
P12	6218.08902	1.510E−23	1.516E−23	0.4	1.534E−23	−1.2	0.1067	0.1006	−5.7
P10	6219.79720	1.380E−23	1.377E−23	−0.2	1.395E−23	−1.3	0.1091	0.1029	5.7
P8	6221.47727	1.180E−23	1.184E−23	0.3	1.199E−23	−1.3	0.1116	0.1058	−5.2
P6	6223.12928	9.410E−24	9.359E−24	−0.5	9.522E−24	−1.7	0.1141	0.1085	−4.9
P4	6224.75324	6.550E−24	6.518E−24	−0.5	6.622E−24	−1.6	0.1190	0.1120	−5.9
P2	6226.34917	3.370E−24	3.357E−24	−0.4	3.403E−24	−1.4	0.1228	0.1170	−4.7

a

Line profile. Data representation in tabular form

Show table | Show form

Show 50 rows starting from 0 | In all rows: 185 | « ‹ › » | Control / Export

State quantum numbers in normal modes representation Nv_i, l_i - initial and f - final state												Intensity I_1 (cm^{-1}/(molecule cm^{-2}) at T=K)	Vacuum wavenumbers (cm^{-1})
v_1^f	v_2^f	l_2^f	v_3^f	r^f	J^f	ε^f	v_1^i	v_2^i	l_2^i	v_3^i	r^i	J^i	ε^i		
2	0	0	1	3	45	e	0	0	0	0	0	46	e	2.32e-23	4813.38892
2	0	0	1	3	43	e	0	0	0	0	0	44	e	3.12e-23	4815.33603
2	0	0	1	3	41	e	0	0	0	0	0	42	e	4.12e-23	4817.26398
2	0	0	1	3	39	e	0	0	0	0	0	40	e	5.35e-23	4819.1713
2	0	0	1	3	37	e	0	0	0	0	0	38	e	6.82e-23	4821.0601
2	0	0	1	3	35	e	0	0	0	0	0	36	e	8.56e-23	4822.93009
2	0	0	1	3	33	e	0	0	0	0	0	34	e	1.05e-22	4824.78154
2	0	0	1	3	31	e	0	0	0	0	0	32	e	1.27e-22	4826.61474
2	0	0	1	3	29	e	0	0	0	0	0	30	e	1.51e-22	4828.42994
2	0	0	1	3	27	e	0	0	0	0	0	28	e	1.76e-22	4830.22738
2	0	0	1	3	25	e	0	0	0	0	0	26	e	2.01e-22	4832.0073
2	0	0	1	3	23	e	0	0	0	0	0	24	e	2.25e-22	4833.76989
2	0	0	1	3	21	e	0	0	0	0	0	22	e	2.46e-22	4835.51536
2	0	0	1	3	19	e	0	0	0	0	0	20	e	2.63e-22	4837.24389
2	0	0	1	3	17	e	0	0	0	0	0	18	e	2.74e-22	4838.95562
2	0	0	1	3	15	e	0	0	0	0	0	16	e	2.78e-22	4840.65071
2	0	0	1	3	13	e	0	0	0	0	0	14	e	2.74e-22	4842.32927
2	0	0	1	3	11	e	0	0	0	0	0	12	e	2.61e-22	4843.99143
2	0	0	1	3	9	e	0	0	0	0	0	10	e	2.37e-22	4845.63726
2	0	0	1	3	7	e	0	0	0	0	0	8	e	2.04e-22	4847.26684
2	0	0	1	3	5	e	0	0	0	0	0	6	e	1.63e-22	4848.88022

b

Annotation on 2017-03-17 11:15:33: Primary public source 2006_ReZePaGr_CO2_T5_CO2_VGT_296K was uploaded by Fazliev Alexander on 2016-05-06 17:15:13

Calculation/Experiment

Substance	
Name	CO_2

Method
Fourier Transform Infrared Spectroscopy (FTIR)

Reference

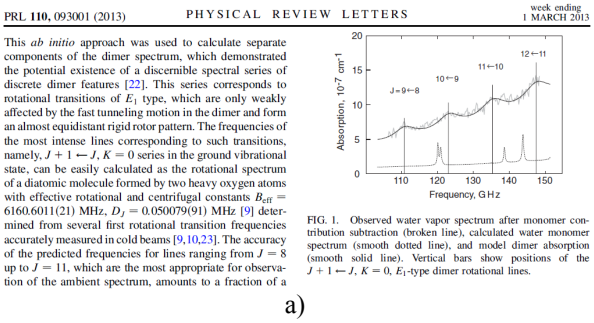
L. Regalia-Jarlot, V. Zeninari, B. Parvitte, A. Grossel, X. Thomas and P. von der Heyden and G. Durry, A complete study of the line intensities of four bands of CO2 around 1.6 and 2.0 mcm: a comparison between Fourier transform and diode laser measurements, Journal of Quantitative Spectroscopy and Radiation Transfer, 2006, 10.1016/j.jqsrt.2005.11.021, Atmospheric carbon dioxide is a key specie for the Earth climate. Two spectral windows at 1.6 μm and 2.0 μm are of particular interest for the in situ and remote monitoring of carbon dioxide from satellite, balloon or airborne platforms using infrared absorption spectroscopy. A precise knowledge of the line strengths is a prerequisite for an accurate concentration retrieval. In this paper, we have revisited in the laboratory the (3001)III←(0000) and (3001)II ← (0000) bands of CO2 near 1.6 μm and the (2001)III ← (0000) and (2001)II ← (0000) bands near 2.0 μm by implementing both a high-resolution Connes-type Fourier-transform spectrometer and a tunable diode laser spectrometer equipped with several telecommunication-type semiconductor laser devices. Approximately 200 (respectively 18) transitions of CO2 have been carefully investigated in spectra recorded with the FT spectrometer (respectively with the tunable diode laser spectrometer). The intensity measurements achieved with both instruments are thoroughly compared to previous instrumental determinations, ab-initio calculations and available atmospheric molecular database.,

Properties of physical quantities (output data)			
Thermodynamic conditions			
Temperature	296 K		
Pressure	1 atm		
Broadening substance			
Broadening substance name	CO_2		
Pressure	0 atm		
Contour type	**Voigt Contour (DL/VGT)**		
Parameter	*Unit*	*Availability*	*Error*
Pressure-induced shift coefficient (δ_L)	cm^{-1} atm^{-1}	-	-
The Doppler profile parameter ($\Delta\sigma_D$)	-	-	-
Pressure-induced broadening coefficient (γ_L)	cm^{-1} atm^{-1}	+	-
Exponential temperature-dependence of shift coefficient (n_δ)	-	-	-
Exponential temperature-dependence of broadening coefficient (n_γ)	-	-	-
Linear temperature-dependence of shift coefficient (n_δ)	-	-	-
Linear temperature-dependence of broadening coefficient (n_γ)	-	-	-

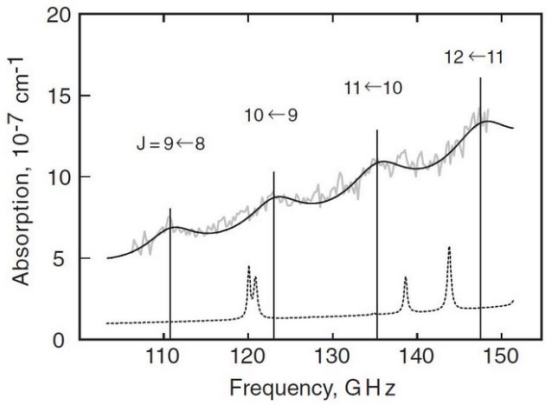
Properties of physical quantities (output data)	
Wavenumbers (ω)	
Unit	cm^{-1}
ω_{min}	4813.38892
ω_{max}	6377.8588
The number of transitions	[185]
Error	-
Intensity	
Unit	cm^{-1}/(molecule cm^{-2})
$(Intensity)_{min}$	9.240e-25
$(Intensity)_{max}$	1.330e-21
Summary intensity	4.241e-20
Availability	+
Error	-
Quantum numbers of transitions	
Quantum number notation	TAbDinfh-1
The number of ro-vibrational bands	
TVAbDinfh-1	[4]
The number of ro-vibrational bands	
TVAbDinfh-1_1	[4]
Total angular momentum (J)	
J_{min}	0
J_{max}	51
Verification of formal and nonformal constraints (including selection rules)	
The number of transitions with unique quantum numbers	[185]
The number of transitions with nonunique quantum numbers	[0]
The number of unassigned transitions	[0]
Results of selection rule verification	
The number of forbidden identifications (Energy levels quantum numbers constraints, l_2=0 & ε≠e)	[0]
The number of forbidden identifications (Nuclear statistics constrains)	[0]
The number of forbidden identifications by (Energy levels quantum numbers constraints, $v_2 \neq l_2 \vee l_2 > J$)	[0]
The number of transitions fail to satisfy the rotational selection rules (ΔJ=0 if e-f) ∨ (ΔJ=±1 if e-e, f-f)	[0]
The number of forbidden transitions (Vibrational selection rules, $\Delta v_2 + \Delta v_3 = 2n$, n=0,1,2,.., Δv=|v'-v''|)	[0]
The number of forbidden transitions (Rotational selection rules, $l_2'=l_2''$=0 - Q - forbidden)	[0]
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)	[185]
The number of transitions that fail to satisfy any constraints	[0]

c

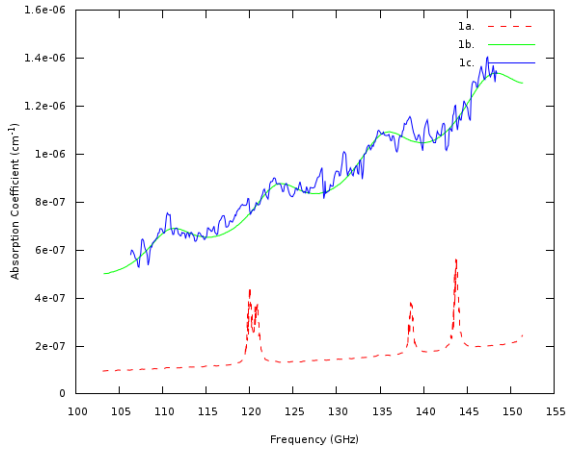
Fig. 1 Models of the paper fragment that contains the tabular representation: (a) source table with measurement data; (b) representation of this table in the information system; and (c) metadata that characterize the properties of the numerical array that is represented in the tables in fragments (a) and (b).



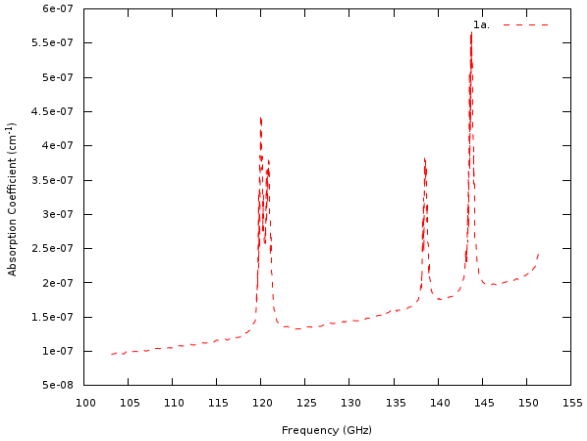
a)



b)



c)



d)

Reference			
M.Yu. Tretyakov,E. A. Serov, M. A. Koshelev, V.V. Parshin, and A. F. Krupnov , Water Dimer Rotationally Resolved Millimeter-Wave Spectrum Observation at Room Temperature, Physical Review Letters, **2013**, **Volume 110**, Issue 9, Pages 093001, DOI: 10.1103/PhysRevLett.110.093001, http://link.aps.org/doi/10.1103/PhysRevLett.110.093001. Annotation			
Axis of abscissa (x-axis) Physical Quantity (Unit)	Frequency (GHz), [Linear]	**Axis of ordinates (y-axis) Physical Quantity (Unit)**	Absorption Coefficient (cm⁻¹), [Linear]
Matter	H_2O	**Physical quantity**	Absorption Coefficient
Temperature	296 K	**Method of measurement (computation)**	Absorption spectroscopy
Pressure	0.0171 atm	**Data type**	Theoretical
Plot title	Figure 1a. Calculated water monomer spectrum		
Modify	Delete		

e)

Fig. 2 Models of a graphical resource of a publication: (a) a fragment of original publication that contains a figure; (b) the figure used for quantization; (c) the complex plot built on the basis of the quantization results of fragment b; (d) an elementary plot from fragment c; and (e) the description of the elementary plot in fragment d.

2.3 Scientific graphics

Scientific plots are used in quantitative spectroscopy fields where exact measurements are lacking in modern experimental techniques (for example, due to the complex atomic composition of a molecule or short-wavelength range), e.g., in the study of continuum absorption important in the investigations of planetary and exoplanetary atmospheres, of spectral properties of weakly bonded molecular complexes and molecules in the UV region necessary for quantitative description of photochemical reactions in the gaseous phase.

Plots with which a user works in the W@DIS IS can be divided into two classes: simple and composite. Simple plots contain only one set of coordinates,

represented by a curve, a set of dots or bars. Composite plots can contain many curves in one coordinate space. There are two types of composite plots in the IS: (1) plots obtained by combining simple plots from one publication and (2) plots obtained from comparison of different data sets from different publications.

A simple plot is a basic data structure in the IS. It is stored as a collection of abscissas and ordinates for the corresponding data set and associated metadata. A set of metadata for each plot includes physical quantities, such as: a substance participating in the physical process described by the plot, the temperature and pressure of the process, the data type (experimental or theoretical), spectral function and method (measurements or calculations), and X- and Y-coordinates and their units of measurement; as well as auxiliary metadata, including: the plot style (a curve representable in several ways or a set of points or bars); linear or logarithmic scales along the abscissa and ordinate, a caption and a commentary for the plot, a bibliographic reference to the paper from which the plot has been taken, and the figure number in this paper. Each simple plot is accompanied by the attached scanned image from the source paper, which allows us to compare the original figure with the plot built automatically in the system. In turn, combining simple plots from one publication, one can obtain a composite plot.

The search and comparison interface allows one to find already loaded plots by a wide range of criteria, such as physical values along the both axes with appropriate units of measurement, substance, temperature, and pressure; or any other physical or auxiliary metadata. As a result of the search, one obtains sets of data from different publications, which can then be combined in one coordinate space for further comparison.

The scientific plots, described in this work, represents the dependencies of physical quantities in 1D–3D Cartesian coordinates. The most common are 2D plots. As a rule, several curves are shown in one plot in one coordinate space, which characterize the behavior of physical parameters under different thermodynamic conditions or provide the comparison of original results by authors with works of other researchers. The number of plots that contains the only curve is relatively small in the total volume of plots published.

The main idea of systematization is a separation of every curve from a set of curves in a complex plot into primitive plots, which is supplemented by a set of metadata describing the plot with the level of detail necessary for searching for it.

Let us give several definitions.

Definition 1. The *primitive plot* is a plot in Cartesian coordinates that contain only one curve from a figure published, in the same coordinate system, relating to the same physical parameter and its measuring units, and a set of metadata describing the plot.

Definition 2. The *composite plot* is a plot in Cartesian coordinates that contain all primitive plots (>1) from a figure published, in the same coordinate system, having the same physical parameters and their measuring units, and sets of metadata describing each plot from this figure.

Definition 3. The *primitive image* in a figure published is an image of one object under study and the related set of metadata that characterizes the properties of the object and its image.

Definition 4. An image that contains more than one primitive image of an object from a figure published is called the *composite image* in the figure published.

In particular, the set of metadata of an elementary image includes a reference to the publication from which the figure described has been extracted. Composite images can be single- or multipaper.

Definition 5. The *primitive figure* is a figure that contains a single scientific plot or image.

Definition 6. The *composite figure* is a figure that contains scientific plots and images.

3 Data and information sources

3.1 Definitions

The variety of molecules for which the problems mentioned in [10] have been solved and the related methods is quite wide. For this reason, solutions to several problems by different methods for different molecules or their isotopologues can be presented in one publication. The solution to one task can be the content of several tables. During systematization of data extracted from publications, such a variety of tables creates many problems, especially in the cases where the solution to a subject task is divided into parts and is represented in several tables. There is no sense to refer individual data arrays to the tables they were extracted from. For this reason, we here use an information object that represents the original data of a publication describing one molecule, one spectroscopy task, and one solution method.

3.1.1 Primitive and composite data sources

This information object shall be called the *data source*. Different data source types are met in scientific papers. Let us give several definitions.

Definition 7. All parts of the published solution to a task of quantitative spectroscopy along with the molecule name, reference, and name of the solution method (or reference to the method description) are called the *"primitive data source"*.

We assume that empty solutions are not published. On the other hand, solutions can include measurement data which go out of date with time or be wrong themselves. A data source the content of which is completely declined by experts is called negligible. The number of such sources in the modern spectroscopy is insignificant.

Definition 8. An information object exhibiting basic properties of a primary source of data cardinality of which differs from unity is called the *composite data source.*

Any expert set of spectral data (e.g., HITRAN [11]) can serve an example of composite data source.

3.1.2 Information source

A primary source can be endowed with additional properties. The list and number of these properties depend on information tasks for solution of which these properties are used. A data source with additional properties is called the source of information.

Definition 9. A *primitive data source* with additional properties is called a *primitive source of information* extracted from a publication.

The source of information is a set of properties and their values attributed to a data source. For a number of information tasks, for example, the search for reliable solutions to quantitative spectroscopy problems, one can select properties values of which are automatically calculated. A source of information usually includes some statements from the publication that contains the data source described by this source of information. The better half of a source of information characterizes the knowledge contained in the publication in an implicit form.

The list of additional properties is determined by a researcher on the basis of information tasks that are to be solved. There are two such tasks in our work: the task of semantic search and the task of automated composition of an expert data set. Let us note that primary sources of information relating to one publication do not contain identical statements. The difference between a publication and a related primary source of information can be significantly smaller than the difference between the publication and a related primary data source. This is due to those additional properties of the task solution in the publication that are included in the definition of a particular source of information. For example, such an additional property can be the description of validity of the solution or the description of the standard deviations of the initial data source from other data sources, etc. In addition, the statements contained in the primary source of information may not be contained in the publication.

3.1.3 Sources of information attributed to pairs of data sources

The representation of a source of information that characterizes the properties of all pairs, including a selected data source with all other data sources, is much more complex. The visualization of such a source of information is necessary for researchers for a number of reasons. First, in spectroscopy, as well as in other data intensive subject domains, it is common to compare the results of experiments performed by different groups. Second, there can be several types of such pair relationships. Third, the number of data sources in the IS varies with time (new works on state and transition parameters appear). Fourth, the measurement accuracy increases; therefore, the values of the criteria that determine the reliability of facts are to be reviewed. Fifth, the number of facts in the comparison between data sources can be tens of thousands, which makes it more convenient to represent them graphically. The representation of this information in the text form is cumbersome and allows one to see only a local picture.

4 Ontology metrics in quantitative spectroscopy

Users of applied data stored in data collections, related to data intensive subject domains, currently meet problems of selection of necessary data, which concern not only the data intension, but also its quality. The ontologically described collections are preferable. Such collections can be objectively compared in terms of metrics of the corresponding ontologies. Naturally, the multiplicity of ontology descriptions gives information about a collection significantly better quality. A certain standard of such a description should arise for each of applied subject domains with time. Below we give an example of the quantitative estimation of the ontology description of resources in the W@DIS IS [12].

As a result of the work, a set of spectral data was collected and systematized within the Molecular Spectroscopy IS for several molecules: H_2O, H_2S, HOCl, OCS, O_3, SO_2, C_2H_2, CH_4, CO_2, CH_3OH, CO, HBr, HCl, HF, HI, N_2, CH_3Br, CH_3Cl, N_2O, NH_3, NO_2, PH_3, and their isotopologues. The numerical array of spectral data in the Molecular Spectroscopy IS is about 80 GB in MySQL database, where most of the data is on H_2O molecule and its isotopologues. The size of the numerical data array could be reduced by the means of additional optimization of the data structure, but then the load on the computing resources of the Molecular Spectroscopy IS would have to significantly increase. To describe the parts of the complete array, the IS contains about 25 GB of metadata stored in the MySQL database, where the overwhelming majority is the quantitative criteria of data quality derived from the calculations of the values of the correlations between pieces of the numerical data. On the basis of the complete 80-GB data array, ontologies of molecular states and transitions are formed, which are represented as XML files in RDF/XML notation of the OWL language of about 280 GB in total size. It should be noted that the OWL language has several syntax notations, from the shortest in the Manchester syntax to the longest in the OWL/XML syntax. The relatively verbose RDF/XML syntax was selected for the representation of OWL ontologies in the Molecular Spectroscopy IS because of historical reasons; this choice seemed optimal in the beginning of the work on ontology representations in the Molecular Spectroscopy IS in 2006.

On the basis of the 25-GB array of metadata, a semantic information model is formed as the ontology of information resources, represented as XML files in the RDF/XML notation of the OWL language of about 3 GB in size. A semantic model of information on spectroscopic graphics in the form of the ontology of spectroscopic plots, represented as an XML file in RDF/XML notation of the OWL language of only 2 MB in size, should be mentioned separately. More complete

quantitative information on resources is given in Table. 1.

The completeness of description of the subject domain and its parts by different applied ontologies is estimated using metrics of the ontologies. Some metrics of the applied ontologies on spectroscopy are given in Table 2.

Table 1. Volume of data, metadata, and ontologies in W@DIS IS

List of resources in W@DIS IS	Volume, GB
Data layer	
Spectral data	80.779
Metadata layer	
Metadata	24.772
Ontology layer	
Ontology of information resources on quantitative spectroscopy	3.231
Ontology of molecular states and transitions	280.079
Ontology of scientific graphics on quantitative spectroscopy	0.002
All resources	398.8

Table 2. Estimation of the metrics of applied ontologies on quantitative spectroscopy

Ontology	Axiom	Logical axiom	Declaration axioms	Class	Object property	Data property	Individual	DL expressivity
OIR	$5.4*10^6$	$4.6*10^6$	606	324	92	355	$1.4*10^6$	ALCHON(D)
OSPM	$0.97*10^9$	$0.9*10^9$	68	30	13	25	$2.0*10^9$	ALC(D)
OSG	$1.81*10^4$	$1.37*10^4$	3690	62	17	10	$3.7*10^3$	ALCHO(D)

OIR means the ontology of information resources, OSPM means the ontology of molecular states and transitions, OSG means the ontology of spectroscopic plots.

5 Conclusion

The aim of the work was focused on ontological description of information resources collections on quantitative spectroscopy. This description give us possibility to organize the semantic search in the domain on the base of traditional criteria of the spectroscopy. The publication models were developed and formalized with help of OWL 2DL. The data and information sources were constructed as a part of the formalization. Description of sources, state, transitions and spectral functions became a basis for the construction of three applied ontologies. These ontologies were used for catalogization of the articles of the quantitative spectroscopy topics and their parts.. The metrics of the ontologies were estimated.

The proposed model can be used under formalization of the information resources of differen type in other subject domains.

Acknowledgments. The work was financially supported by the Russian Foundation for Basic Research (grant no. 07-13-0411).

References

[1] Tim Berners-Lee, James Hendler and Ora Lassilla, The Semantic Web, Scientific American, May 17, 2001.

[2] L. Kalinichenko, A. Fazliev, E. Gordov, N. Kiselyova, D. Kovaleva, O. Malkov, I. kladnikov, N. Podkolodny, N. Ponomareva, A. Pozanenko, S. Stupnikov, A. Volnova, New Data Access Challenges for Data Intensive Research in Russia, CEUR Workshop Proceedings, v. 1536, 2015, P.215-237, 17-th International Conference on Data Analytics and Management in Data Intensive Domains, DAMDID/RCDL 2015; Obninsk; Russian Federation; 13 - 16 October 2015; Code 118237.

[3] Keller-Rudek, H., Moortgat, G. K., Sander, R., and Sörensen, R., The MPI-Mainz UV/VIS spectral atlas of gaseous molecules of atmospheric interest, Earth System Science Data, 5, 365–373, (2013) doi:10.5281/zenodo.6951.

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

[5] N. A. Lavrentiev, O. B. Rodimova, A. Z. Fazliev, A. A. Vigasin, "Systematization of published

research graphics characterizing weakly bound molecular complexes with carbon dioxide," Proc. SPIE 10466, 23rd International Symposium on Atmospheric and Ocean Optics: Atmospheric Physics, 104660E (30 November 2017); doi: 10.1117/12.2289932.

[6] N.A. Lavrentyev, M.M. Makogon, A.Z. Fazliev, Comparison of the HITRAN and GEISA Spectral Databases Taking into Account the Restriction on Publication of Spectral Data, Atmospheric and Oceanic Optics, 2011, Vol. 24, No. 5, pp. 436–451.

[7] A.Privezentsev, D.Tsarkov, A.Fazliev, J.Tennyson, 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, http://ceur-ws.org/Vol-614/ owled2010_submission_6.pdf.

[8] S. S. Voronina, A. I. Privezentsev, D V. Tsarkov, A. Z. Fazliev, An Ontological Description of States and Transitions in Quantitative Spectroscopy, Proc. of SPIE XX-th International Symposium on Atmospheric and Ocean Optics: Atmospheric Physics, 2014, Vol. 9292, 92920C.

[9] N. A. Lavrentiev, O. B. Rodimova, A. Z. Fazlie v, Systematization of graphically plotted published spectral functions of weakly bound water complexes, Proc. SPIE of 22nd International Symposium Atmospheric and Ocean Optics: Atmospheric Physics, Eds. Gennadii G. Matvienko; Oleg A. Romanovskii, Tomsk, Russian Federation, v. 10035, 100350C (November 29, 2016); doi: 10.1117/12.2249159.

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

[11] 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, Volume 130, Pages 4-50, DOI: 10.1016/j.jqsrt.2013.07.002.

[12] A. Akhlyostin, Z. Apanovich, A. Fazliev, A. Kozodoev, N. Lavrentiev, A. Privezentsev, O. Rodimova, S. Voronina, A.G. Csaszar, J. Tennyson, The current status of the W@DIS information system, Proc. SPIE of 22-nd International Symposium Atmospheric and Ocean Optics: Atmospheric Physics, Eds. Gennadii G. Matvienko; Oleg A. Romanovskii, Tomsk, Russian Federation, v. 10035, 100350D (November 29, 2016); doi: 10.1117/12.2249235.