Computed Knowledge Base for Quantitative Spectroscopy

Privezentsev A.I.[1], Tsarkov D.V.[2], and Fazliev A.Z.[1]

[1] V.E.Zuev Institute of Atmospheric Optics SB RAS, Tomsk, Russia
[2] The University of Manchester, Manchester, UK

Abstract. A knowledge base for an ontological search and integration and systematization of information resources for quantitative molecular spectroscopy is described. The knowledge base is presented in the form of applied ontologies used for solving the foregoing tasks. A data model needed for realization of the information system in quantitative spectroscopy is developed. We have modeled the subject domain in question on the basis of a set of publications from different journals on molecular spectroscopy. The properties inherent to the tasks associated with quantitative spectroscopy are defined. To this end, use is made of the OWL DL language. Information sources characterizing the properties of the tasks under consideration enable solutions of spectroscopic tasks to be classified and a semantic search for valid data to be organized.

Keywords: Knowledge Base, Quantitative Spectroscopy, Applied Ontology

1 Introduction

Major components of a research into an analysis of subject domains are as follows: gathering facts, developing domain models, comparing an original model with those of other researchers, providing access to the model for other researchers, and, finally, publishing the models. This brings up the following questions: Is the set of facts collected and generated by the researcher complete? Are these facts consistent with each other? Does the formal language of the model specification allow one to build a domain model adequate to the facts gathered? Is the original model consistent with those developed by other researchers? How can quick access to the results of investigations be provided?, etc.

These topical questions also arise in molecular spectroscopy which is one of the fields of physics widely used in many applied research areas. Note that molecular spectroscopy is known to study the spectral properties of molecules. Spectral data sets require systematization of facts and development of software for data processing. Software implementation implies construction of domain models related to these data as well as development of the required facilities associated with a search for information resources. In spectroscopy, such resources are solutions of spectroscopic tasks. In quantitative spectroscopy, different research groups have gathered sets of expert data [10, 5, 7]. These sets are found

to disagree with each other. Reconciling the data available with those obtained by different investigators involved in collecting expert data is one of the main tasks not only in spectroscopy, but also in other subject areas.

In the mid-2000s, an IUPAC project was launched [11, 12] wherein a task was undertaken to develop an information system intended for collection of all currently available data on water molecules and water isotopologues and containing facilities for identification of contradictions between data on vacuum wavenumbers. There are formal and informal contradictory data reconciliation criteria. The former criteria include selection rules, root-mean-square values, the difference between identical transition vacuum wavenumbers, etc. The latter criteria imply expert estimates of data quality like those used in the information system discussed here. The ontology of information resources for water molecules was built to describe the properties of published solutions of spectroscopic tasks [3]. At present the ontology contains description of all information resources on molecules of sulfur dioxide, ammonia, phosphine, carbon oxide and dioxide, and methane. The ontology of information resources for spectroscopy developed in the W@DIS information system provides a description of the state-of-the art of the published data in this subject area and can be accessed via the Internet (`http://wadis.saga.iao.ru`).

The initial stage in the development of the ABox of the ontology is data import in the information system. Most of the individuals used in our ontology corresponding to the properties of an imported data source are formed in the import process. One individual corresponds to a set of properties of its associated data source, whereas other individuals characterize pair relations between this data source and other data sources in the information system. Users can update the selected ontology by adding a certain class, which extension satisfies their information query. Such classes may contain their associated subclasses. The problem statement and problem solution involve a lot of details discussed in [8, 6, 9, 1].

2 Data and knowledge models in quantitative spectroscopy

In order to develop a software realization of an ontological search for spectroscopic information resources, a definition must be given of domain models related to the resources. In our earlier work [8, 6], we defined three domains: "Matter", "Quantitative Spectroscopy", and "Information Source". The first two domains are closely related because quantitative spectroscopy studies the spectral properties of matter. The latter properties, in their turn, are closely related to different processes of matter – radiation interaction, atomic and molecular collisions, etc.

2.1 "Matter" domain

There are quite a few descriptive models of the "Matter" domain. We will examine only part of this domain related to the data model under consideration. It is pertinent to note that only isolated atoms and molecules are formalized,

whereas the substances in different phase states are omitted. The software for describing matter in this approximation was developed in the framework of the ATMOS project [4].

2.2 "Quantitative Spectroscopy" domain

The "Quantitative Molecular Spectroscopy" domain describes spectral molecular properties typical for emission and absorption in the atmosphere of planets. A special feature of our model [3] is that it is based on published primary data sources containing solutions of direct and inverse tasks characterizing physical quantities related to emission and absorption processes. There is a number of tasks dealt with in molecular spectroscopy. These tasks can be divided into several groups. Since the discussion is restricted to the emission and absorption properties we are interested in a group of tasks related to measurements or calculations of the spectral line parameters required for describing the processes under study. These tasks form a structure consisting of two chains [2]: a chain of direct tasks and that of inverse tasks. An XML scheme used to solve this kind of tasks was put forward in [8].

2.3 "Information Source" domain

A great body of information is acquired from published articles written in natural languages. As this takes place, a technical task is to build subject-predicative structures, where the basic element is a statement. The data source and information source are more complicated objects. Being part of a publication, a data source contains a solution of a spectroscopic task. An information source contains the solution properties related to the corresponding data source.

A researcher defines a list of additional properties based on the information tasks to be solved. In the present work, only one of the tasks of this kind — an ontological search — is discussed. Note that primary information sources related to one publication do not contain identical statements. The difference between a publication and a primary information source and a data source integrated into a single object may be negligibly small as compared to the difference between a publication and a primary data source. This difference is due to additional properties of the solution of the task included in the definition of one or another information source. For example, the additional properties may include root-mean-square deviations of the primary data source from other data sources, etc. What is more, the statements contained in the primary information source may be absent in the publication to which it is related.

3 Information resources ontology for molecular spectroscopy

The basis for construction of an ontology for spectroscopic tasks is an approach wherein the task is an object described by means of an input-processing-output model. Thus, among the metadata included in the consideration of input and

output data are their associated intension and a number of attributes describing the quantitative properties of the data extensions. In the model chosen, the metadata for input data are references to an URI. The quantitative characteristics contained in the metadata are generated in a dynamic way in uploading user files that include solutions of spectroscopic tasks.

The spectroscopic ontology consists of three different layers. The top layer corresponds to a base ontology that defines the basic entities from the spectroscopic domain. For example, the part corresponding to quantitative spectroscopy contains classes referred to as State and Process. The intermediate layer corresponds to the application ontology that describes the classes and properties of molecules in quantum theory and input and output data for spectroscopic tasks. The bottom layer contains end-user ontologies that can be build using W@DIS information system and/or other methods.

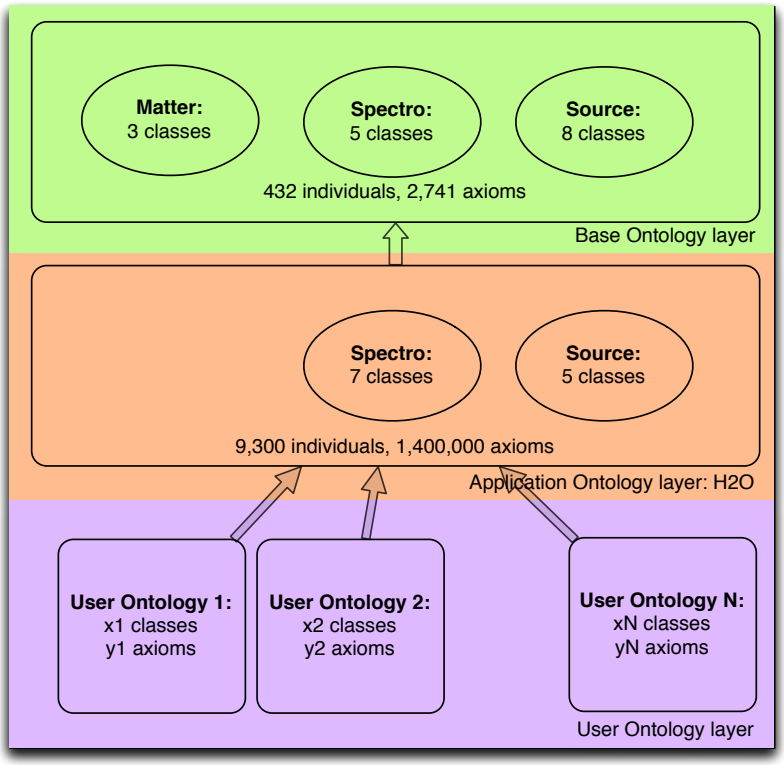

Fig. 1. Structure of information resources of molecular spectroscopy

Figure 1 shows a layered structure together with some statistics for classes, individuals, and axioms used in the base and application ontologies. The statistics for the application ontology corresponds to a water molecule together with

Table 1. The number of individuals in the ontology for water and carbon dioxide molecules.

Task	Water		Carbon dioxide	
	{I1}	{I2}	{I1}	{I2}
T2, T6	140/2*105	2543/3.6*105	124/0.8*105	998/0.6*105
T3, T5	357/2.1*105	6088/5.5*105	338/5.5*105	1537/3.1*105

water isotopologies. Similar data for a carbon dioxide molecule are presented in Table 1.

3.1 The properties defined in the base and application ontologies

The base ontology defines 14 object properties and 4 datatype properties. The application ontology defines 9 object properties and 14 datatype properties. The actual number of datatype properties is much larger (see below), but it depends on the type of the quantum number representation.

3.2 Individuals

In any ontology, individuals correspond to particular domains. The "Matter" domain contains individuals for different atoms and molecules and types of phase states. The "Quantitative Spectroscopy" domain contains individuals for tasks (T1–T6) and types of the quantum number representations. The individuals of the "Information Source" domain can be divided into two parts. One part {I1} describes properties inherent to solutions of the tasks, whereas the other part {I2} defines the properties of pairs of data sources being solutions of the tasks. A property of a pair of data sources is, e.g., the root-mean-square deviation or a maximum difference between vacuum wave numbers of the same transitions in the pair.

The numerators in the table correspond to the number of individuals in a given pair of tasks. The denominators specify the number of axioms (property assertions) of these individuals. Every individual corresponding to task {I1} possesses a fixed number of property assertions (e.g., for a water molecule and tasks T2 and T3, individuals contain 40 and 61 assertions, respectively). The number of axioms corresponding to an individual from {I2} is not fixed. For example, for the root-mean-square deviation, the number of property assertions is $131+5N_{bands}$, where N_{bands} is the number of bands with the same vibrational quantum numbers.

4 Applications

The ABoxes of the base and application ontologies for spectroscopic information resources contain individuals each of which represents properties of its associated data source or those of a pair of data sources. The set of classes defined in these

ontologies corresponds to a set of desired queries. Users may construct classes that satisfy a restriction on properties in accordance with their intentions relating to an analysis of spectroscopic data. The database in question incorporates classes corresponding to findings of the ontological search.

Below is a list of classes corresponding to plausible user queries.

1. Data sources containing a single vibrational band;
2. All data sources including the root-mean-square deviation for selected vibrational bands;
3. *Data sources containing transitions in a selected range of wavenumbers;*
4. Data sources that satisfy a set of properties related to formal constraints;
5. Data sources with no transitions rejected by experts;
6. Data sources containing the same transitions as in the data source;
7. Data sources containing only unique transitions;
8. Canonical information sources;
9. Pairs of correlated information sources containing only measured data;
10. Pairs of correlated information sources containing only a selected vibrational band;
11. *Vibrational bands found in direct tasks.*

Let us use two classes to illustrate the mechanism of an ontological search for information resources. To understand the reasoning behind the formulation of the information tasks, let us consider an illustrative example. Planning investigations into the spectral characteristics of molecules, we must be aware of the results of the earlier work on the subject. In particular, calculated and measured transition parameters can be compared provided that relevant observation data on certain vibrational bands are available. On the other hand, further experiments cannot be designed without knowledge of particular vibrational bands where measurements have already been performed. In the application ontology of information resources a VibrationalBand class contains instances representing sets of measured identical vibrational band properties. Note that results obtained have been published. As vibrational bands are characteristics of transitions, we have to deal with solutions of direct T2, T3 (calculations) and inverse tasks T5, T6 (measurements).

Assume the following query: Which vibrational bands of the $H_2^{17}O$ molecule are known to be calculated? This query is encoded as a class named "VIBRATIONAL BANDS FOUND IN DIRECT TASKS". This is a subclass of the "VIBRATIONALBAND" class. This class is defined as a set of restrictions on properties whose definition (using the Manchester syntax) is as follows:

VIBRATIONALBAND **and inverse** HASQUANTUMNUMBERSOFBAND **some**
(**inverse** HASVIBRATIONALBAND_MD **some**
(**inverse** HASTRANSITIONQUANTUMNUMBERS_MD **some**
(**inverse** HASOUTPUTDATA_MD **some**
(INFORMATIONSOURCE **and**
(ISSOLUTIONOF **value** T2 **or**
ISSOLUTIONOF **value** T3))))).

To form a set of measured (as opposed to calculated) vibrational bands, it suffices to replace direct tasks (T2,T3) with inverse tasks (T5,T6) in the restrictions. Note that for this molecule, the number of calculated vibrational bands is 12354, and the number of measured ones is 63.

Another example is a typical search query that looks for information sources. It contains data in a given range of vacuum wavenumbers. The resulting publications correspond to the same spectroscopic tasks as in the previous example. The result of the query is an extension of the class named "DATA SOURCES CONTAINING TRANSITIONS IN A SELECTED RANGE OF WAVENUMBERS". The extension is defined by means of the following property restrictions (written in the Manchester syntax):

(T2-IS **or** T3-IS **or** T5-IS **or** T6-IS) **that**
HASOUTPUTDATA_MD **some** (HASWAVENUMBERS_MD **some**
(WAVENUMBERS_MD **and**
((HASMINWAVENUMBER **some float**[$>= 0.0, < 10.0$]) **or**
(HASMAXWAVENUMBER **some float**[$>= 0.0, < 10.0$])))).

In this example, the wavenumber range corresponds to microwaves. The extension of this class for the $H_2^{17}O$ molecule contains information about 22 data sources. Note that the numbers used in restrictions are query parameters defined by users according to the information task at hand.

The domain experts can add other classes to this list. Untrained users may employ the list leaving it unchanged. Experienced users can regard the queries as patterns and adapt them to their own needs by changing the values of the properties. Advanced users can create new classes using all possible ways of constructing queries.

5 Summary

A computed ontology of information resources for quantitative spectroscopy is presented. The ontology contains base and application ontologies. In the information system containing these ontologies users can update the ontologies to include other classes created by means of imposing restrictions on some properties. To perform the task, users can employ a set of classes available in the information system or use the classes as patterns for creating new classes or those unrelated to the patterns.

The ontology of information sources for quantitative spectroscopy contains a set of ontologies related to certain molecules and solutions of tasks involving a definition of transition property values as well as molecular states.

Work is under way on the development of ontologies of information resources for atmospheric chemistry and atmospheric radiation. The ontology development method described in this paper shows good promise for these applications. Moreover, the atmospheric chemistry resources are closely related to those pertaining to molecular spectroscopy. This relation is realized due to the use of absorption cross sections in the description of photochemical reactions.

The work was supported by RFBR (Grant 11-07-00660).

References

1. Akhlyostin, A., Kozodoev, A., Lavrentiev, N., Privezentsev, A., Fazliev, A.: Computed knowledge base for description of information resourses of molecular spectroscopy 4. Software. Russian Digital Library Journal 15(3) (2012), http://www.elbib.ru/index.phtml?page=elbib/eng/journal/2012/part3/AKLPF
2. Bykov, A., Fazliev, A., Kozodoev, A., Privezentsev, A., Sinitsa, L., Tonkov, M., Filippov, N., Tretyakov, M.: Distributed information system on molecular spectroscopy. In: Proceedings of SPIE. vol. 6580, p. 65800W (2006)
3. Fazliev, A., Privezentsev, A., Tsarkov, D., Tennyson, J.: Computed Knowledge Base for Description of Information Resources of Water Spectroscopy. In: Sirin, E., Clark, K. (eds.) OWLED. CEUR Workshop Proceedings, vol. 614. CEUR-WS.org (2010)
4. Gordov, E., Lykosov, V., Fazliev, A.: Web portal on environmental sciences "ATMOS". Advances in Geosciences 8, 33–38 (2006)
5. Jacquinet-Husson, N., Scott, N., Chédin, A., Crépeau, L., Armante, R., Capelle, V., Orphal, J., Coustenis, A., Boonne, C., Poulet-Crovisier, N., et al.: The GEISA spectroscopic database: Current and future archive for Earth and planetary atmosphere studies. Journal of Quantitative Spectroscopy and Radiative Transfer 109(6), 1043–1059 (2008)
6. Lavrentiev, N., Privezentsev, A., Fazliev, A.: Computed knowledge base for description of information resourses of molecular spectroscopy 2. Data model of quantitative spectroscopy. Russian Digital Library Journal 14(2) (2011), http://www.elbib.ru/index.phtml?page=elbib/eng/journal/2011/part2/LPF
7. Pickett, H., Poynter, R., Cohen, E., Delitsky, M., Pearson, J., Muller, H.: Submillimeter, millimeter, and microwave spectral line catalog. Journal of Quantitative Spectroscopy and Radiative Transfer 60(5), 883–890 (1998)
8. Privezentsev, A., Fazliev, A.: Computed knowledge base for description of information resourses of molecular spectroscopy 1. Basic concepts. Russian Digital Library Journal 14(1) (2011), http://www.elbib.ru/index.phtml?page=elbib/eng/journal/2011/part1/PF
9. Privezentsev, A., Tsarkov, D., Fazliev, A.: Computed knowledge base for description of information resourses of molecular spectroscopy 3. Basic and applied ontologies. Russian Digital Library Journal 15(2) (2012), http://www.elbib.ru/index.phtml?page=elbib/eng/journal/2012/part2/PTF
10. Rothman, L., Gordon, I., Barbe, A., Benner, D., Bernath, P., Birk, M., Boudon, V., Brown, L., Campargue, A., Champion, J., et al.: The HITRAN 2008 molecular spectroscopic database. Journal of Quantitative Spectroscopy and Radiative Transfer 110(9-10), 533–572 (2009)
11. Tennyson, J., Bernath, P., Brown, L., Campargue, A., Carleer, M., Császár, A., Gamache, R., Hodges, J., Jenouvrier, A., Naumenko, O., et al.: IUPAC critical evaluation of the rotational-vibrational spectra of water vapor. Part I–Energy levels and transition wavenumbers for $H_2^{17}O$ and $H_2^{18}O$. Journal of Quantitative Spectroscopy and Radiative Transfer 110(9-10), 573–596 (2009)
12. Tennyson, J., Bernath, P., Brown, L., Campargue, A., Császár, A., Daumont, L., Gamache, R., Hodges, J., Naumenko, O., Polyansky, O., et al.: IUPAC critical evaluation of the rotational-vibrational spectra of water vapor. Part II: Energy levels and transition wavenumbers for $HD^{16}O$, $HD^{17}O$, and $HD^{18}O$. Journal of Quantitative Spectroscopy and Radiative Transfer 111(15), 2160–2184 (2010)