Computed Knowledge Base for Description of Information Resources of Water Spectroscopy

Alexander Fazliev[1], Alexey Privezentsev[1], Dmitry Tsarkov[2], and Jonathan Tennyson[3]

[1] Institute of Atmospheric Optics SB RAS, Zuev Square. 1, 634021 Tomsk, Russia
[2] University of Manchester, Oxford Road, Manchester M13 9PL, UK
[3] University College London, Gover St. London WC1E 6BU, UK
{Alexander Fazliev, Alexey Privesetsev, faz, remake}@iao.ru
Dmitry Tsarkov, tsarkov@cs.man.ac.uk
Jonathan Tennyson, jtennyson@ucl.ac.uk

Abstract. We develop the addition to the W@DIS information system that allows one to load solutions of the quantitative spectroscopy problems. These solutions are supplied with calculated semantic annotations that characterise properties of the solutions. In addition to the typical properties (e.g. represented in Dublin Core) the solution reliability properties are also determined. Every solution is represented in a knowledge base as an individual of the quantitative spectroscopy ontology, which contains more that 10^6 axioms. In the paper we present the knowledge base structure and describe two classes of the information sources classification tasks, together with the solutions using querying OWL ontologies.

Keywords: Water Spectroscopy Knowledge Base, Annotation Model, Classification Problems.

1 Introduction

The interest to the spectral properties of the water molecule is caused by its exceptional status. Water participates in many processes on Earth, including the life processes. This molecule is one of the most studied ones. In this paper we present a knowledge base which is capable to describe the full set of the spectral properties of the water molecule on a logical level. This work requires the united effort of three groups of professionals in spectroscopy, information systems and logics.

In the last decade molecular spectroscopists have united to address key tasks requiring a cooperative solution. One of these tasks is the development of a comprehensive representation of the spectrum of the water molecule. A suitable theoretical strategy for representing the spectrum was formulated in the framework of two projects [1-3]. The first protocol, “Marvel”, is generally applicable for molecular spectroscopy [1]. A collective effort from domain specialists was made for collecting and validating both calculated and measured spectral data [2].

In order to collect and represent spectroscopic data an information system W@DIS (http://wadis.saga.iao.ru) with three layers architecture [4] was implemented. In this

architecture a knowledge layer contains scientific annotations of the spectroscopic problem solutions [5]. These annotations were represented as individuals of OWL-ontologies in W@DIS.

As a result of this joint work, the knowledge base describing properties of water spectroscopy problems' solutions was created.

To solve the classification problems of quantitative water spectroscopy we use the Description Logic reasoner FaCT++ [6,7]. The size of data in the knowledge base restrict the authors in the complexity of the TBox part of the ontology and force them to adjust the ABox structure according to the facilities of the reasoning system.

2 Molecular Spectroscopy Model

In this work we consider the conceptualizations of two domains. The concepts related to the description of physical system states are included into spectroscopy model with high level of granularity, while the concepts that characterize the processes of transition from one state into another are not described in details. *De facto* such domain model is widely used, for example, in a series of physics domains in which the procedures of experimental results acquisition are well-established but the amount of data is huge requiring many years experimental measurements. One of such domains, i.e. water quantitative spectroscopy, is described in the paper.

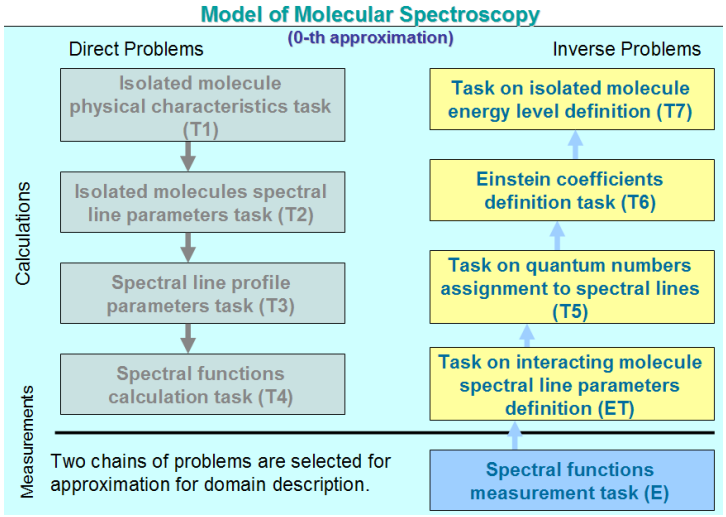

Fig. 1. Molecular quantitative spectroscopy model.

The simplification of procedure knowledge is caused by the assumption due to the following fact: in practice, for example, in information description of quantitative spectroscopy the most needed information is the quantitative information on molecules states. In quantitative spectroscopy the procedure knowledge may be of

interest only to a narrow circle of specialists that implement different methods of domain problems' solution.

Schematic model of quantitative molecular spectroscopy for the information sources of which the scientific annotations are composed automatically is presented in Fig. 1.

In the model of a domain the solutions of the molecular spectroscopy problem were considered as domain data. Every solution of the certain problem has annotation that is a set of metadata. This set contains properties of solution of a spectroscopic problem for a definite molecule in a certain physical conditions published in a journal, a monograph or in the Internet together with the values of these properties. We name such metadata as an *annotation* of information resource.

3 The Knowledge Base Structure

The knowledge base (KB) consists of the two parts, each represented as a set of OWL ontologies. Fig. 2 shows the KB structure, where arrows corresponds to the import statements.

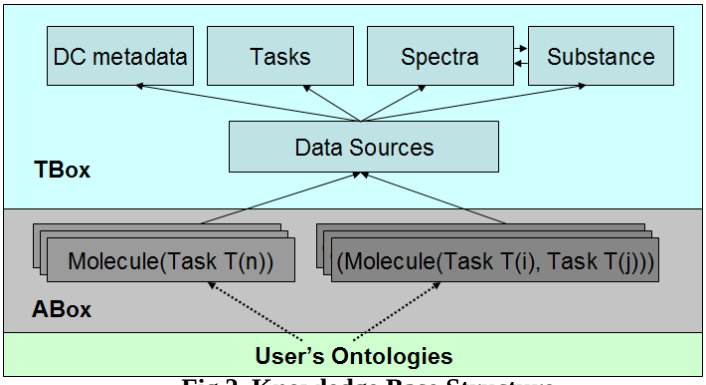

Fig.2. Knowledge Base Structure

The TBox contains five basic ontologies that describe the spectroscopy domain. The *DataSources* ontology describes papers on spectroscopy. It also describes properties of the solutions of the water spectroscopy problems that are published this way. The *Spectra* ontology describes the spectrum characteristics of the matter, the *Substance* one describes molecules, etc.

Overall TBox contains 75 classes, 41 object properties and 127 data properties. It also contains 63 individuals that represent methods to solve spectroscopy problems, physical measurement units, the spectral line form and other physical quantities. The expressivity of a TBox is ALCHOIN(D).

The ABox of a KB contains more that a million facts that are obtained from the IS. These facts were generated by IS application during the analysis of the spectroscopy problems' solutions. The ABox is split into two parts. One part describes properties of the solutions of the water spectroscopy problems, based on one article (about 55k axioms in 62 files). Another part describes the standard deviation between solutions of the same problems from different papers, and contains more that 1,300,000 axioms in 39 files.

In W@DIS information system a user can add new facts independently of the others. Thus it is possible to use this feature to create ontologies for that user's specific tasks. Some of such tasks are described below.

The KB currently contains the description of properties of solutions for the water molecules isotopomers (Molecule= H_2O, $H_2{}^{17}O$, $H_2{}^{18}O$, HDO, $HD^{17}O$, $HD^{18}O$, D_2O). The parameter *n* (see Fig.1) takes values 1,2,3,5,6,7, and a pair (i,j) takes values (1,7), (2,6), (3,5). All components of the KB are publicly available and can be downloaded from the IS W@DIS website (http://wadis.saga.iao.ru/saga2/ontology).

4 Two Classification Problems

The Water Spectroscopy KB allows one to classify the physical quantities according to their properties. Here we describe two such problems.

The first problem is to classify all the published information sources as *reliable* and *unreliable*. There are several such classifications as there are several validity criteria. Table 1 illustrates the results of reliable information sources selection from the analysis of about 800 scientific annotations. In this table the *m(n)* line for a given water isotopomer and a problem T means that out of *m* publications that contains a solution for the problem T only *n* contains reliable solutions.

Task /Molecule	Calculations of energy levels (T1)	Measurement of energy levels (T7)	Calculations of wavenumbers (T2)	Measurement of wavenumbers (T6)
H_2O	9 (2)	30 (24)	5 (0)	91 (47)
$H_2{}^{17}O$*	4 (0)	19 (15)	5 (1)	40 (31)
$H_2{}^{18}O$*	4 (0)	18 (18)	5 (1)	59 (35)
HDO	1 (0)	32 (28)	3 (0)	83 (56)
$HD^{17}O$		3 (3)	2 (0)	3 (3)
$HD^{18}O$		5 (4)	2 (0)	6 (6)
D_2O	1 (1)	18 (8)	3 (0)	38 (26)
$D_2{}^{17}O$			1 (0)	3 (3)
$D_2{}^{18}O$			2 (0)	3 (3)
Total	15 (3)	125 (100)	28 (2)	318 (207)

Table 1. The results of checking the selection rules in primary data sources [5]

The validity was checked according to the restrictions on the values of the quantum number that came from selection rules for transitions and restrictions on rotational quantum numbers for energy levels. The asterisk symbol indicates that an annotation for a molecule was modified according to the comments of experts on quantum number assignment correction. It is easy to see that the solutions of T1 and T2 problems contain a bigger percentage of data arrays that in turn contain solutions with incorrect quantum numbers. An example of a query that describes a restriction on the quantum numbers values is shown in Section 5.

The second problem is to classify values obtained from the water spectroscopy problems' solutions by means of root-mean-square (standard) deviations. It can be viewed as a refinement of the previous problem. Some quantum numbers assigned to a certain physical quantity can satisfy the validity criteria but the assignment itself can be incorrect. The value of a standard deviation helps to figure out whether the values were calculated inaccurately or incorrect assignment was made.

5 Query Examples

The reasoning over an ontology is used to answer queries about its elements. The typical queries are used to determine properties of a solution of a spectroscopy problem (see Fig. 1). These queries are built by restricting values of certain properties. As an example lets create a query to find all the canonical information sources for water isotopomer $H_2{}^{17}O$ and problems T1 and T7 (in the *canonical* information source quantum numbers satisfies all chosen criteria, described as a selection rules):

InformationSource **that** hasSubstance **value** H2_17O
    and hasOutputData_MD **some** (hasTransitionQuantumNumbers_MD **some**
(hasQuantumNumbersType **value** NormalModes
and hasNumberOfNonuniqueTransitions **some** {0}
and (hasNumberOfUnlabeledTransitions **some** {0}
    or hasNumberOfUnassignedTransitions **some** {0})
and hasNumberOfInvalidIdentifications **some** {0}
and hasNumberOfInvalidTransitions **some** {0}
and hasNumberOfInvalidWaterTransitions **some** {0}
and hasNumberOfInvalidWater-C2V-Transitions **some** {0}
and hasNumberOfRejectedTransitions **some** {0}))

The query structure depends on the type of the problems (here T1 and T7) and the symmetry group of the molecule (here C_{2v}), that define the number of selection rules. Using similar queries allows one to separate information sources to canonical and non-canonical. Note that some domains have weaker notions of canonicity so the query can be simplified.

Another problem of the information source classification can be solved using the restrictions on the values of standard deviation of some physical quantity. Here is an example of query (PublicationsWithLargeDeviation_Band_0_5_1_0_5_0) which is used to find all publications in which the value of a standard deviation for the vibration band 051-050 is over 0.1 cm^{-1} (note that most of the details related to spectroscopy physics are omitted here):

isRMSMemberOf **some** (hasRMSBandPair **some** (hasRMSTransitionBand **some**
(hasTransitionQuantumNumbersOfBand **value** Identification_on_NormalModes_0_5_1_0_
5_0_Band
and hasBandRMSDeviationValue **some** float[> 0.1])))

This query can be useful for a person that require some physical parameters related to given vibration band in the following way. If the query corresponding to class PublicationsWithLargeDeviation_Band_0_5_1_0_5_0 is non-empty (i.e., contains some information sources) then these sources are unreliable w.r.t. given band. In this case the additional check of the values of corresponding physical characteristics is necessary.

6 Conclusions

In this paper we present a model of quantitative molecular spectroscopy. We describe an information system W@DIS that contains data about molecular spectroscopy problems' solutions from the published papers. In particular, it contains the complete set of facts about solutions for the water spectroscopy published in the last 60 years.

Using this model as a knowledge domain we develop a knowledge base that describes properties of the solutions of the common problems in the domain area. We describe two classes of problems of information sources classification that can be solved in the KB. We provide examples of queries to the KB that describes classes in the OWL ontology that solve the necessary problems.

As a part of the future work we plan to update the KB with facts about other important molecules, including methane, carbon dioxide, carbon oxide, acetylene, ammonia, and others.

References

1. T. Furtenbacher, A.G. Császár and J. Tennyson, MARVEL: measured active rotational-vibrational energy levels, J. Molec. Spectrosc., v 245, 2007, p. 115-125.
2. J.Tennyson, P.F.Bernath, L.R.Brown, *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 $H_2{}^{18}O$, J. Quant. Spectr. Rad. Transfer, 2009, v. 110, Pages 573-596.
3. IUPAC project N 2004-035-1-100 «A database of water transitions from experiment and theory». http://www.iupac.org/web/ins/2004-035-1-100.
4. De Roure D., Jennings N., Shadbolt N. A Future *e*-Science Infrastructure // Report commissioned for EPSRC/DTI Core e-Science Programme. 2001. 78 p.
5. Privesentsev A.I., Ontological knowledge base implementation and software for information resources description in molecular spectroscopy, Tomsk State University, PhD Dissertation, 2009, 238 Pages.
6. Dmitry Tsarkov and Ian Horrocks. *FaCT++ Description Logic Reasoner: System Description.* In Proc. of the Int. Joint Conf. on Automated Reasoning (IJCAR 2006), volume 4130 of Lecture Notes in Artificial Intelligence, pages 292-297. Springer, 2006.
7. Dmitry Tsarkov, Ian Horrocks, and Peter F. Patel-Schneider. *Optimizing Terminological Reasoning for Expressive Description Logics.* J. of Automated Reasoning, 39(3):277-316, 2007.