VAMDC XML-schema is initial point in the Semantic Web. Will Trust be final point?

Fazliev A.Z.[1], Privezentsev A.I.[1], Tsarkov D.V.[2], and Tennyson J.[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. Department of Physics and Astronomy, University College London, London WC1E 6BTUK

An interoperable infrastructure for spectral data transport to users and user query formation for choosing information of interest has been built within a VAMDC project. A domain data model that provides a basis for the infrastructure is characterized by a VAMDC XML-schema describing species, processes, environment, and bibliographic references. A graphical user interface makes it possible to formulate queries relating to species, processes, and environment. The infrastructure in question enables researchers to address expert data from different providers, whereas providers, in their turn, can identify and remove inconsistencies available in expert data at their disposal.

The VAMDC project is nearing completion. This raises the questions as to further development of the infrastructure and particular information technologies to be used. The answers to these questions are offered and discussed.

When dealing with spectral expert data, we could see that in many instances, they contain misprints and/or information whose validity is questionable.

This is why it is believed that the infrastructure needs to be supplemented with an XML-schema describing results of a detailed analysis of the data quality. This is the answer to the question about further development of the infrastructure. The answer to the question concerning information technologies is not so unambiguous. This will largely depend on professional qualifications of the staff of providers maintaining information resources. In this work, use is made of Semantic Web technologies. Why do the users need information on the data quality? In applied sciences, spectral data are part of input data for solving tasks of relevance to one or another research. The sensitivity of solutions of the applied tasks to the quality of spectral data varies over a wide range. There are different definitions of the term “data quality”.

Since there is no consensus among experts and investigators regarding the meaning of this term, this circumstance may result in incorrect solutions of the tasks at hand (e.g., in the vicinity of bifurcation points) not only in a quantitative sense, but from a qualitative standpoint as well.

A task characterized by high sensitivity to input data in molecular spectroscopy is exemplified by computations of reference energy levels. For isotopologues of the water molecule and that of hydrogen sulfide, this task was solved in [1, 2] and in [3], respectively. In solving the task in question, we have built information system W@DIS to provide data collection, storage, and quality analysis in accordance with certain formal criteria. The data acquired from published sources were uploaded into W@DIS, and annotations on the data were generated automatically. The annotations were represented in OWL 2 DL and corresponded to individuals of the ontology of the information resources of molecular spectroscopy. A major advantage of the OWL 2 DL language, besides the fact that it is a W3C standard recommendation, is the … support of reasoning in this language using very efficient inference engines, like FaCT++ [5]. In addition to this, the expressive power of OWL 2 DL, unlike that of the other OWL 2 profile, is enough to express the core properties of the data quality problem in the quantitative spectroscopy domain. One of such problems is to compute the agreement between properties of multisets of transitions and states for different molecules.

This ontological knowledge base on the properties of multisets and its associated inference engine make up an expert system known as a computer system that emulates a logical decision-making ability of a human expert. A system of this kind consists of three parts: an inference engine, a knowledge base and a graphical user interface.

The knowledge base incorporates the properties of data acquired from publications. In the database, information relating to molecules is stored in three tables summarizing

1. The characteristics of states of an isolated molecule;
2. The characteristics of transitions of an isolated molecule;
3. The parameters of spectral lines of molecular gases.

The database is a basis for generating different lists of spectral characteristics of the methane molecule. The lists are used for developing expert datasets of spectral line parameters, computation of reference energy levels, etc.

The central problems with the use of the lists lie in the incorrectness and inconsistency of some of the data included in the lists.

To incorporate 'The data quality analysis' domain in the VAMDC XML-schema, the key tasks for providers are to agree on a list of methods for a data quality analysis and to represent the results of the analysis in the XML-schema form or in terms of the OWL DL ontology. The ontologies developed here can be used in solving this task.

1. 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. Spectrosc. Radiat. Transfer, 2009, V. 110, No.9, P. 573-596.
2. J. Tennyson, P. F. Bernath, L.R. Brown, at 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 $HD^{18}O$, J. Quant. Spectrosc. Radiat. Transfer, 2010, V. 111, No.15, P. 2160-2184.
3. E.R.Polovtseva, N.A.Lavrentiev, S.S.Voronina, et al., Information System for Molecular Spectroscopy. 5. Ro-vibrational Transitions and Energy Levels of the Hydrogen Sulfide Molecule, Atmospheric and Oceanic Optics, 2012, Vol. 25, No. 2, pp. 157–165.
4. A.Fazliev, A.Privezentsev, D.Tsarkov, and 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
5. D.Tsarkov, I.Horrocks. FaCT++ Description Logic Reasoner: System Description. Automated Reasoning, Third International Joint Conference, IJCAR 2006, Seattle, WA, USA, August 17-20, 2006, Lecture Notes in Computer Science 4130, Springer 2006.