Geophysical Research Abstracts,
Vol. 11, EGU2009-10857, 2009
EGU General Assembly 2009

{Data and metadata schemes in chains of direct and inverse problems in spectroscopy of water}

A.Fazliev (1), A.Kozodoev (1), N.Lavrentiev (1), A.Privezentsev (1), J.Tennyson (2)

(1) Institute of Atmospheric Optics SB RAS, Tomsk, Russia, (2) University College London, London, UK

(faz@iao.ru /Fax: +7 3822 492086 / Phone: 7 3822 492277)

A. Fazliev (1) and the A.Fazliev Team

(1) Institute of atmospheric optics SB RAS, Integrated information systems center, Tomsk, Russian Federation (faz@iao.ru), (2) University College London, London, UK

An idea to develop procedure knowledge domain model in a form of task net in information system has been proposed. Tasks solutions are interpreted as data. Solution properties are regarded as metadata. Water spectroscopy is a knowledge domain in which a good approximation for task net would be a pair of chains of direct and inverse tasks. In such an approximation data schemes are the basis of knowledge domain conceptualization. Data scheme represents the next level of water spectroscopy representation granulation. The work describes metadata and data schemes for eight tasks of molecular spectroscopy.

The importance of results of water spectroscopy is great. Precise and valid information on water is necessary in many applied knowledge domains such as atmospheric optics, astronomy, atmospheric radiation and so on. The report describes metadata and data layer in W@DIS information system oriented on information representation. An important feature of the ICS is its spectral data validity check realized in the explicit form.

The main sets of molecules spectral characteristics available to consumers have been formed in the last forty years. These are such data banks as HITRAN[1], GEISA[2] and others. Data on spectral line parameters and interfaces for their operation appeared for the first time in the Internet in "Atmospheric gases spectroscopy"[3] information system. In the above works this data representation in a form of files and interfaces for their operation hasn't solve the main problem (in our opinion) of spectral data in the information systems. This is the problem of creation of accessible applications developed to check the validity of data gathered in an information system. One of the components necessary for automatic data validity check is the presence of computer processable initial results of measurements and calculations. Bibliographic references that can simplify the solution of this task are present in the explicit form in data files presented by Hitran and GEISA. However, they lack initial data related to these references.

Gathering of heterogeneous data obtained by different groups of spectroscopists becomes more sensible if these groups cooperate. In this case, such problems as format agreement, knowledge domain conceptualizations agreement, ontologies formation agreement and etc... are solved basing on the agreement of these groups. In Russia this step was made in 2005 (RFBR grant "Distributed information system "Molecular Spectroscopy"") when organizations from Tomsk, Nizhnii Novgorod and Saint-Petersburg united in order to create distributed information system for molecular spectroscopy. The results of the work done in the framework of this project are published in Ref. [4−6]. IUPAC project[7] on systematization of data on water and its isotopomers spectral parameters was being realized at the same time. Problems of data validity were studied very thoroughly in the framework of this project. As a result of cooperation of participants of these projects the formulation of problem of information structures unification in spectroscopy was done. "Information source" concept is the basis of the solution of this task. In the framework of

the agreements reached information sources structure, methods of development, storage, processing and delivery were formalized. As a result *W@DIS* information system[9] oriented on the solution of the above problem on the example of water molecule has been created.

This report describes the stage of cooperated work dedicated to the creation of information sources and data upload. In the framework of Semantic Grid [8] paradigm for information system creation, its three-layer model has been selected. It consists of data and calculations layer, information layer and knowledge layer. In this article we describe only the data layer. The details of description of information layer and information system (on water spectroscopy) knowledge layer are presented in Ref. [5,6].

1 INFORMATION RESOURCES STRUCTURE

Data schemes used in domains are mainly determined by its conceptualization. In procedure knowledge domain a conceptualization depends on its tasks and determines its granulation. That is why developing molecular spectroscopy information system we selected a model of two chains of tasks. In the framework of this model only those tasks solution results have sense that are either uploaded into ICS or calculated in it and then uploaded into knowledge database. In our work uploading, delivery, storage and representation of solution results (task output data) are the main goals.

The chains we selected contain eight tasks[4]. Each of them has its own data scheme composed of data intensions. Input data and some values of qualitative and quantitative properties of task solution results form metadata. For example, in inverse task of molecule energy levels determination - energy level, quantum numbers (in three notations), uncertainty in determination of energy level and number of transitions used for energy level determination are the intensions.

Data upload task solved in the creation of *W@DIS* ICS (http://wadis.saga.iao.ru) implied:

- creation of information sources formation means
- creation of data schemes formation means and
- creation of formation means for data validity check for such constraints as compliance with data types and rules that determine permitted values and relations between data intension values (first of all, quantum numbers)

2 DATA UPLOAD SYSTEM

In the selected model of molecular spectroscopy the total number of entities (intensions) and relations for two chains of molecular spectroscopy tasks exceeds two hundred. Molecular spectroscopy conceptualization assumes that energy level, which is a solution of a direct task, is not equivalent to energy level obtained in the solution of inverse task, i.e. their values can be different even if the values of other compared attributes are equal.

There is a constraint on files uploaded into ICS. Each of them should be a text file of columnar structure with fixed columns width and should contain a set of nonrepeatable entities in a row. All the lines in this file should be of identical structure.

In the given system data upload and adjusting for the organization of results of solution of a definite task of a chain in domain database is actualized by application programmer. An XML-scheme is created for each task. This scheme describes the structure of uploaded data as well as some constraints on the values. A list of molecular spectroscopy concepts was created by a knowledge engineer. Basing on this list data scheme of each task was formulated. A menu that defines a sequence of user operations in the process of data input was created. The scripts (PHP) that implement these operations were developed.

REFERENCES

1. L.S. Rothman, D. Jacquemart, A. Barbe, D.Chris Benner, M. Birk, L.R. Brown, M.R. Carleer, C. Chackerian, Jr, K. Chance, V. Dana, V.M. Devi, J.-M. Flaud, R.R. Gamache, A. Goldman, J.-M. Hartmann, K.W. Jucks, A.G. Maki, J.-Y. Mandin, S.T. Massie, J. Orphal, A. Perrin, C.P. Rinsland, M.A.H. Smith, J. Tennyson, R.N. Tolchenov, R.A. Toth, J. Vander Auwera, P. Varanasi, G. Wagner, The *HITRAN* 2004 Molecular

Spectroscopic Database, Journal of Quantitative Spectroscopy & Radiative Transfer 96 (2005) 139–204, Data bank HITRAN, http://cfa-www.harvard.edu/hitran/

2. Geisa,
   http://ether.ipsl.jussieu.fr/

3. Internet-accessible information system "Spectroscopy of Atmospheric Gases",
   http://spectra.iao.ru

4. A.D.Bykov, A.Z. Fazliev, N.N.Filippov, A.V. Kozodoev, A.I.Privezentsev, L.N.Sinitsa, M.V.Tonkov and M.Yu.Tretyakov, Distributed information system on atmospheric spectroscopy, Geophysical Research Abstracts, SRef-ID: 1607-7962/gra/EGU2007-A-01906, 2007, v. 9, p. 01906.

5. A.V. Kozodoev. A.I.Prevezentsev, A.Z. Fazliev Annotation of information resources in "Molecular spectroscopy" distributed information system, Electronic Libraries, 2006, Ch.9, ver.3 (in Russian) http://www.elbib.ru/index.phtml?page=elbib/rus/journal/2006/part3/ KPF

6. A.I.Prevezentsev, A.Z. Fazliev Applied task ontology for molecular spectroscopy information resources systematization. The Proceedings of 9^{th} Russian scientific conference "Electronic Libraries: Advanced methods and technologies, electronic collections" - RCDL'2007, Pereslavl Zalesskii, 2007, part.1, 2007, P.201-210.

7. IUPAC project No.2004-035-1-100 "A database of water transitions from experiment and theory"
   http://www.iupac.org/web/ins/2004-035-1-100

8. De Roure D., Jennings N., Shadbolt N., A Future e-Science Infrastructure, Report commissioned for EPSRC/DTI Core e-Science Programme, 2001, 78p.

9. A.Z.Fazliev, A.G.Csaszar, J.Tennyson, *W@DIS*: Water spectroscopy with a Distributed Information System // Proc. of the 10 HITRAN Database Conference, 2008, p.38-39.