Information system for molecular spectroscopy. 2. Array operations for transformation of data on spectral line parameters

A.V. Kozodoev and A.Z. Fazliev

Institute of Atmospheric Optics,
Siberian Branch of the Russian Academy of Sciences, Tomsk

Received July 18, 2005

The HITRAN and GEISA databanks use a line structure of data for storing the spectral line parameters. This structure is convenient for use in applications, but it is practically ineffective in storing data in an information system and is inapplicable for computer data processing. In this paper, we propose a modified structure of data on spectral line parameters. The modification implies the addition of three attributes to physical quantities and the values of the spectral line widths broadened by a buffer gas as well as the temperature dependence of these widths. The list of buffer gases includes water vapor, argon, neon, carbon dioxide, and so on. The modified structure of the data is used for collective processing of these data. A set of operations is proposed for use in compilation of the compound databank. The implementation of the operations is discussed for processing the spectral line parameters stored.

Introduction

Development of molecular spectroscopy today is accompanied by accumulation of huge data bulks, first of all, on spectral line parameters. Because the experiments and calculations in such studies are too laborious the research laboratories restrict their efforts to studies of only some molecules. As a rule, even studying a very limited number of molecules, they cover just a part of spectral line parameters. Thus obtained data are published as separate articles.

In practice, electronic publications of the applications containing numerical data on the parameters studied have become more and more popular now. Most of such data are presented in the format adopted in the HITRAN databank.[1] Such an approach has considerably simplified data dissemination. The fact that many older programs employ this format is explained by traditions and some features of the programming languages that were earlier used in computational applications.

Typical programs for spectroscopic computations that are often supplied by the information systems[2–4] as service made us to change our understanding of the requirements to structuring the databanks of spectral line parameters. There are two reasons for changing the format structure suggested by HITRAN authors:

first, this format does not assume spread-outs (it does not suppose input of new physical quantities, for example, half-widths of spectral lines due to broadening by inert gases);

second, machine processing of the attributes of physical parameter in it needs for much extra efforts.

Also, we would like to emphasize the necessity of inputting metadata while working with spectral line parameters. We think that metadata, which describe data operations play a special role here. And this type of metadata scripts[5] is just beginning to appear.

Our approach to structuring the spectral line parameters is aimed at facilitating machine exchange of such sets and introducing these sets to information systems. It allows one to eliminate overheads that result, for instance, from work with different versions of the HITRAN and GEISA databanks. Note that spectral data in such databanks are updated once in two or three years. Therefore, we find it reasonable to create the tools that would help collecting and arranging complex databanks on spectral line parameters available on the Internet.

In the first section of the article, we determine the elemental structure of the databank and the attributes of physical quantities that belong to parameters of a spectral line. We list restrictions in the databank formation.

In the second section, we describe operations that are necessary to form a databank. We emphasize the importance of storage in the information-computational system (ICS) of the databanks with original data sets. Restrictions to original datasets and binary operations are also mentioned.

The fourth section is devoted to realization of these operations within the SQL queries for the spectral line parameters that are stored in the relational database.

The program modules being currently developed for working with spectral line parameters are a part of the information system on the atmospheric spectroscopy (http://saga.atmos.iao.ru).

0235-6880/05/09 680-05 $02.00

1. Description of elements of a databank of spectral line parameters

Parameters of spectral lines are the most demanded data in calculations of both spectral functions and atmospheric radiation characteristics. Parameters of spectral lines are arranged in the form of databanks. Databank structure is shown in Fig. 1.

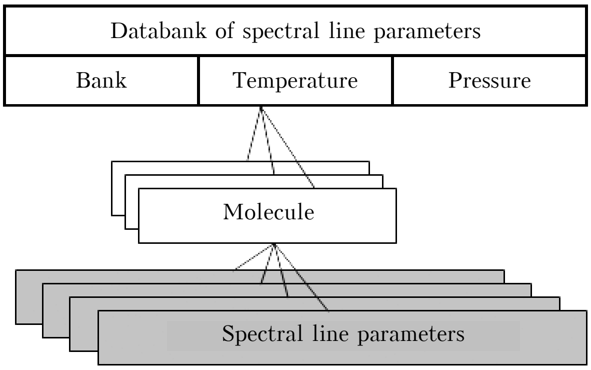

Fig. 1. The structure of databank of spectral line parameters.[1]

Databanks such as HITRAN are stored as files, where each file contains a number of rows of the same structure. Row content is name and code of a molecule, numerical data on some physical parameters and their attributes, and a set of symbols to identify the particular spectral line. The use of this databank structure for machine exchange of data between information systems is technically complicated, but adopting a new structure of parameter bank can lift the problem.

Before we start description of the new method of databank formation, let us consider in detail the bottom text row in Fig. 1. Table 1 gives the physical quantities and their attributes belonging to spectral line parameters. The table italicizes the attributes that we think can be added to the standard (currently standard) list of quantities and attributes, which are, in their turn, printed in bold type. The "+" sign means that the physical quantity possesses a certain attribute.

Error value in the format adopted in HITRAN is described by the class number. In publications, they often use the absolute values of measurement errors in a physical quantity. In order to account this, we introduce the attribute *error type*, which can be either *absolute* or *interval*. Therefore, error value can be either an integer or a real number. For qualitative understanding of the nature of obtaining the value of a physical quantity, we have introduced the attribute *value type*, which varies between *experiment* and *calculation*. To allow input of data that contain experimental values of line intensity in relative units, we anticipated the attribute *scale*, whose "values" are *absolute* and *relative*.

The ICS on atmospheric spectroscopy employs the following list of buffer gases: air, water vapor, molecular oxygen, molecular nitrogen, argon, helium, neon, and carbon dioxide. In a standard bank of spectral line parameters, only air is included as a buffer gas.

Figure 2 shows a detailed structure of the databank of spectral line parameters with all the above suggestions. This structure is a tree and is described by the XML data markup language.[6] The XML-scheme that we suggest for description of this structure is available on the Internet.[7]

Now, consider the restrictions on the spectral line parameters in the databank. No other spectral line with the same identification can be met in the bank of spectral line parameters for any molecule. Names of a bank and its related attributes, temperature, and pressure are applied to all the spectral line parameters in this databank.

Table 1. List of physical quantities forming the spectral line parameter set and their attributes

Physical quantity	**Notation**	**Bibliographic link (b)**	**Error value (e)**	*Error type (t)*	*Value type (v)*	*Scale (s)*
Identification	ID					
Transition frequency	TF	+	+	+	+	
Line intensity	LI	+	+	+	+	+
Buffer gas-**broadened width**	HW	+	+	+	+	
Self-broadened width	HWS	+	+	+	+	
Pressure shift	SH				+	
Temperature dependence for *buffer gas*-**broadened width**	TD				+	
Einstein coefficient	CE					
Lower State Energy	LSE					
Lower statistical weight	LSW					
Upper statistical weight	USW					
Flag for mixing	FM					

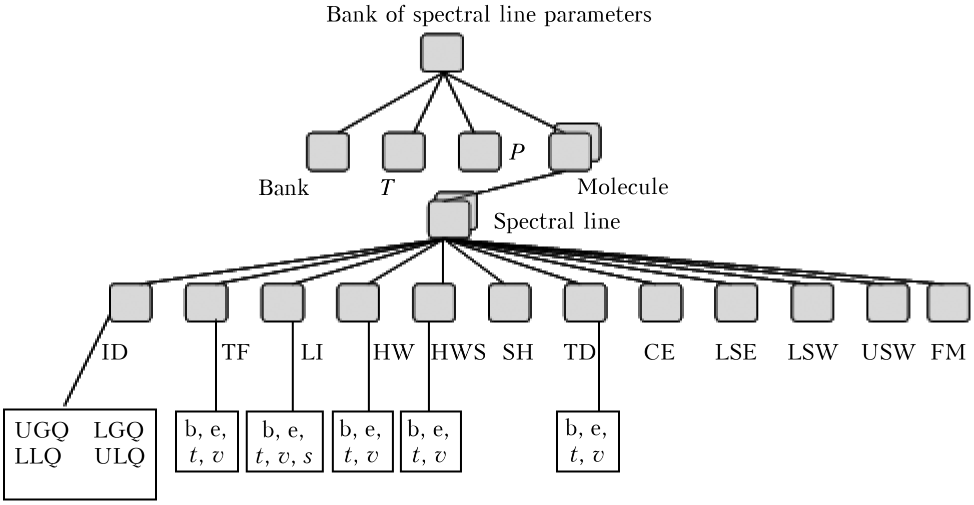

Fig. 2. Detailed structure of the bank of spectral line parameters.

2. Operations of transformation to original datasets

Let a dataset (of spectral line parameters) be the data with the structure of a spectral line parameter bank, but with no values of temperature and pressure. Let the original dataset be a set, where the attribute *Bibliographic link* has the same value for all the physical quantities throughout the dataset, and where the attributes *value type*, *error type*, and *scale* have constant values for a certain physical quantity. The data input system in the ICS on atmospheric spectroscopy allows creation of only original datasets (banks). Operations on datasets (databanks) allow the user to form complex databanks. Operations on a data row (or rows) are a key mechanism in formation of the complex data sources.

In this work, the operations of dataset transformations are those, which can change data structure, but do not touch the values of physical quantities and their attributes. We distinguish between two types of operations: unary and binary.

Unary operations involve selection of rows by numerical ratios (=, <, >) for physical quantities (for example, selection of weak and strong lines), selection by an attribute symbol (e.g., all the rows with the values of experimental intensity), and selection using the attributes connected with identification of spectral line (selection of line or spectral band). We also added to these operations, that of reduction of the number of physical quantities in a row (except for identification, transition frequency, and the line intensity).

Binary operations with data rows include:

1) selection of lines having the same vibrational-rotational identification in a couple of datasets (dataset **intersection**);

2) selections from two sets of spectral line parameters of those identified lines that do not have a pair in identification (dataset **complement**);

3) combination of datasets that have no rows with the same identification (**integration** of datasets);

4) catenation of those datasets, where for every identified line from one set of spectral line parameters there is a corresponding identified line from another set of spectral line parameters (operations of **catenation** for a pair of rows having the same identification).

Note that all catenation operations are non-commutative.

Results of integration and catenation for two simple data sets shown in Fig. 3 are given in Figs. 4 and 5. All the abbreviations used in Figs. 3–5 are explained in Table 1.

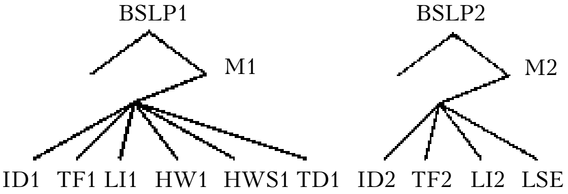

Fig. 3. Original elementary data sets.

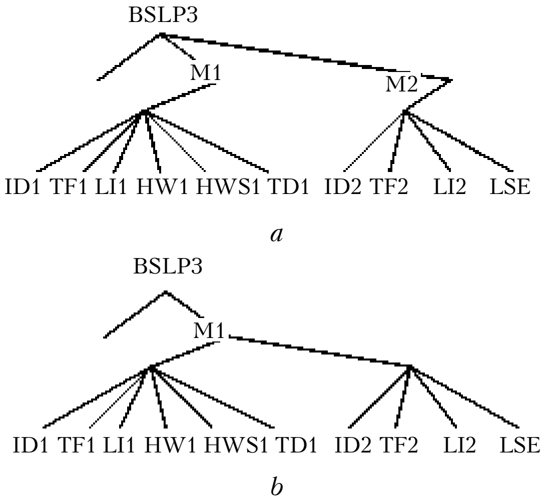
BSLP3
M1
M2
ID1 TF1 LI1 HW1 HWS1 TD1 ID2 TF2 LI2 LSE
a
BSLP3
M1
ID1 TF1 LI1 HW1 HWS1 TD1 ID2 TF2 LI2 LSE
b

Fig. 4. Sample results of applied integration: M1 ≠ M2 (*a*); M1 = M2, ID1 ≠ ID2 (*b*).

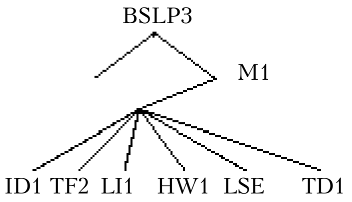

Fig. 5. Example of catenation: M1 = M2, ID1 = ID2.

3. Data storage

Inconveniences of the fixed format of a HITRAN text file can be eliminated by use of XML data formatting [www.w3c.com]. In this case, we can speak about file structure rather than file format. The changes of the structure are seldom and minor. Thus, changes in representation of numbers or in the number of meaningful signs do not influence the methods of working with the file. File structure is described by the XML-scheme, a link to it can be placed in the file.

Data storage in the XML-format allows one to have a good data structure and description (using the XML-scheme). In other words, XML-formatting is good for data dissemination as an alternative to a text file, which has a fixed format. A drawback of such a data representation is an extremely ineffective search of XML-files, because this format is first of all intended for data storage.

Structured data storage and the results of a competitive search can be obtained by means of Database Management Systems (DBMS). The DBMS data structuring potential is comparable to that of the XML-formatting.

A DBMS is quite suitable tool for data access and storage. It allows one to perform storage of structured data, effective data search, selection, and other operations on stored data.

4. Realization of data operations in relational algebra

Modern DBMS offer multiple tools for working with data. One of them is SQL query language. This language is convenient, since it allows many important operations to be performed by its own means only, with no need for resorting to additional tools.

Consider realization of operations using the DBMS supported by the SQL. In description of operations, we shall use the notions adopted for the relational model.[8]

Unary operations

Selection determines the resulting relation. The latter will contain only those tuples, which meet a specified requirement.

Projection determines the new relation, which will contain a vertical subset of the original relation and consist in fetching the values of desired attributes and removing the duplications.

These operations can be fully performed with the SELECT statement of the SQL query language. Syntax of this statement allows these operations to be performed separately as well as jointly.

Binary operations

Integration is concatenation of all the tuples from original relations to a single relation with the duplications removed. A new relation will have the same power, and an equal or lower cardinality. As to spectral line parameters, the original relations must not overlap in either spectral range or vibration-rotational identification of the spectral lines.

This operation is performed using the SELECT and UNION statements, which allow one to make desired selections and unify them, respectively. It is just the resulting selection that is added to the database.

By catenation, we mean combining by equivalence of attributes. The relation power will be equal to the sum of powers of the original relations with subtraction of the number of attributes used for combining. A non-commutative catenation is also possible with the left or right catenation open, when one operand gives out all the tuples, and another one gives only those that meet the catenation requirements.

In the case of spectral line parameters, catenation can only be performed for the vibrational-rotational identification of spectral lines. A catenation in the form described above can be used only to scan data, but never to input it to the database, because in the database, there is only one tuple attribute (one column in a row) per each spectral line parameter. To meet this requirement, we need to perform such a projection over the obtained catenation that each of the resulting tuples has one attribute corresponding to every spectral line parameter. If we address the above example (see Fig. 5), the rules applied to projection can be found in Table 2. A plus sign in a column means that the value of a certain physical quantity is assumed by the corresponding physical quantity in the resulting data structure.

Table 2. Rules for making a projection

Dataset	I	TF	HWS	SH	TD	LSE
1	+	+			+	
2			+	+		+

Because of restrictions to the set of spectral line parameters, we obtain the following: from the viewpoint of relational algebra, operations of catenation described in Section 2 consist of catenation and projection.

To perform these operations in the SQL language, SELECT statement is used together with JOIN.

Conclusion

In this work, the idea of expanding the list of physical quantities used as spectral line parameters in HITRAN databank is considered. The expanded data structure is thought to be applied to online data exchange between information systems. Operations on the sets of spectral line parameters are described. Application of these operations to the data in relational databases is suggested.

The introduced operations after their formal RDF-description can be used as metadata. Such metadata characterize the processes of formation of complex data sources.

Acknowledgments

The authors would like to acknowledge the Russian Foundation for Basic Research for financial support (Grant No. 05–07–90196).

References

1. L.S. Rothman, D. Jacquemart, A. Barbe, D. Chris Benner, M. Birk, L.R. Brown, M.R. Carleerf, C. Chackerian, K. Chancea Jr., 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, and G. Wagner, *The HITRAN 2004 Molecular Spectroscopic Database,* http://www.hitran.com.

2. Yu.L. Babikov, A. Barbe, V.F. Golovko, and Vl.G. Tyuterev, in: *Proc. of 3rd All-Russian Conference on Electronic Libraries* (Petrozavodsk, 2001), pp. 183–187, *Spectroscopy of Atmospheric Gases*, http://spectra.iao.ru.

3. S. Mikhailenko, Yu. Babikov, Vl.G. Tyuterev, and A. Barbe, Computat. Technol. **7** (special issue), 64–70 (2002), *Spectroscopy and Molecular Properties of Ozone*, http://ozone.iao.ru.

4. A.D. Bykov, B.A. Voronin, A.V. Kozodoev, N.A. Lavrentiev, O.B. Rodimova, and A.Z. Fazliev, Atmos. Oceanic Opt. **17**, No. 11, 816–820 (2004), *Atmospheric Spectroscopy*, http://saga.atmos.iao.ru.

5. J.F. Sowa, *Knowledge Representation: Logical, Philosophical, and Computational Foundations* (Brooks Cole Publishing Co., Pacific Grove, CA, 2000), 594 pp., *Top-Level Categories*, http://www.jfsowa.com/ontology/toplevel.htm.

6. Extensible Markup Language (XML) 1.0 (Third Edition), http://www.w3.org/TR/2003/PER-xml-20031030.

7. XML-scheme for description of spectral line parameter databank, http://saga.atmos.iao.ru/data/xsd/SpectralLines-3.xsd.

8. T. Konnoli, K. Berg, and A. Strachan, *Databases: Designing, Realization, and Maintenance. Theory and Practice* [Russian translation] (Vilyams, Moscow, 2000), 1120 pp.