ISSN 0005-1055, Automatic Documentation and Mathematical Linguistics, 2018, Vol. 52, No. 6, pp. 297–305. © Allerton Press, Inc., 2018.
Original Russian Text © S.V. Trepalin, Yu.E. Bessonov, B.S. Fel'dman, E.V. Kochetova, N.I. Churakova, L.M. Koroleva, 2018, published in Nauchno-Tekhnicheskaya Informatsiya, Seriya 2: Informatsionnye Protsessy i Sistemy, 2018, No. 11, pp. 23–31.

The Structural Chemical Database of the All-Russian Institute for Scientific and Technical Information, Russian Academy of Sciences. An Autonomous System for Structural Searches

S. V. Trepalin[a, *], Yu. E. Bessonov[b], B. S. Fel'dman[b], E. V. Kochetova[b], N. I. Churakova[b], and L. M. Koroleva[b]

[a]Institute of Physiologically Active Compounds, Russian Academy of Sciences, Moscow, 142432 Russia
[b]All-Russian Institute for Scientific and Technical Information, Russian Academy of Sciences, Moscow, 125190 Russia
**e-mail: trep@chemical-block.com*

Received July 2, 2018

Abstract—A system is described that provides autonomous searching for information by various characteristics of chemical compounds in the Structural Chemical Database of the All-Russian Institute for Scientific and Technical Information of the Russian Academy of Sciences. Examples of a search and the prospects for further development of the system are given.

Keywords: structural chemical database, information retrieval systems for chemical information

DOI: 10.3103/S0005105518060055

INTRODUCTION

Specialized databases of structural chemical information are an integral part of the information-support system for chemical specialists. Understanding the mechanism of chemical processes and the development of the synthesis of new materials are impossible without the use of such information. According to some data [1], information about the structure and properties of chemical compounds makes up approximately 85% of the total flow of chemical information.

Since 1975, the All-Russian Institute for Scientific and Technical Information of the Russian Academy of Sciences (VINITI RAS) has been work on creating a specialized database of structural chemical information (SDB), the largest in Russia and the third largest repository of structural chemical data in the world. At present, the SDB contains information on more than 7 million chemical structures, approximately 4 million chemical reactions, and 15 million properties of chemical compounds. The array of information accumulated since 1975 includes multi-format data on the structure of chemical compounds, their physicochemical properties, chemical reactions, methods of synthesis, analysis, biological activity, toxicity, and fields of application in chemistry, agriculture, pharmacology, medicine, metallurgy, electrical engineering, physics, ecology, and other parameters. The SDB consists of two components: the base of the structures of chemical compounds and the base of reactions in which these compounds participate. The database is now filled using the CBASE32 software complex, which is a development of the CBASE shell [2]. This software complex provides input and processing of structural, factual and bibliographic data. Graphic information about chemical compounds and reactions is accompanied by various textual information presented in the SDB in the form of subject characteristics (terms), whose system was developed in the VINITI and is currently constantly updated [3]. The subject characteristics are various hierarchically listed information about the compound coded in capital letters of the Latin alphabet: physicochemical properties of the compound, characteristics of its production, its activity, fields of application, etc. Thus, the SDB contains relevant carefully selected information critically evaluated and processed by highly qualified chemical specialists, which is one of the major advantages of the SDB among the diverse flow of chemical information provided to users in the contemporary information space.

The indisputable advantage of any database is the availability of various ways to search for the information the user needs. Each database has its own search tool. The software-technological complex of the SDB includes software systems that use two methods of providing information to users: online searching and autonomous searching. The online search system was described in detail in [4]. This article discusses an autonomous system for structural data searches that makes it possible to search for information in an offline mode in a local database of a system formed from a database of structural chemical information.

TECHNICAL REQUIREMENTS TO THE AUTONOMOUS SEARCH SYSTEM

A modern search system should have a comfortable intuitive interface and the ability to quickly relevant search through all elements of the database information array. During the development of the search system of the SDB, the following requirements were identified:

- a user-friendly graphical interface;
- a quick search with visual results;
- searching using such characteristics of chemical compounds as:

—the molecular formula;

—a name or a fragment of a name;

—an exact structure and structural fragment;

—subject characteristics;

- searching for additional information on the Internet;
- display of bibliographic data from the VINITI RAS catalog;
- display of the following information in search results:

—the name of the chemical compound and its structural formula;

—bibliographic references;

—the characteristics of the subject with comments;

—data on chemical reactions in which the chemical compound participates.

THE DATA CONVERTER. THE BASIC PRINCIPLES OF SEARCH ORGANIZATION

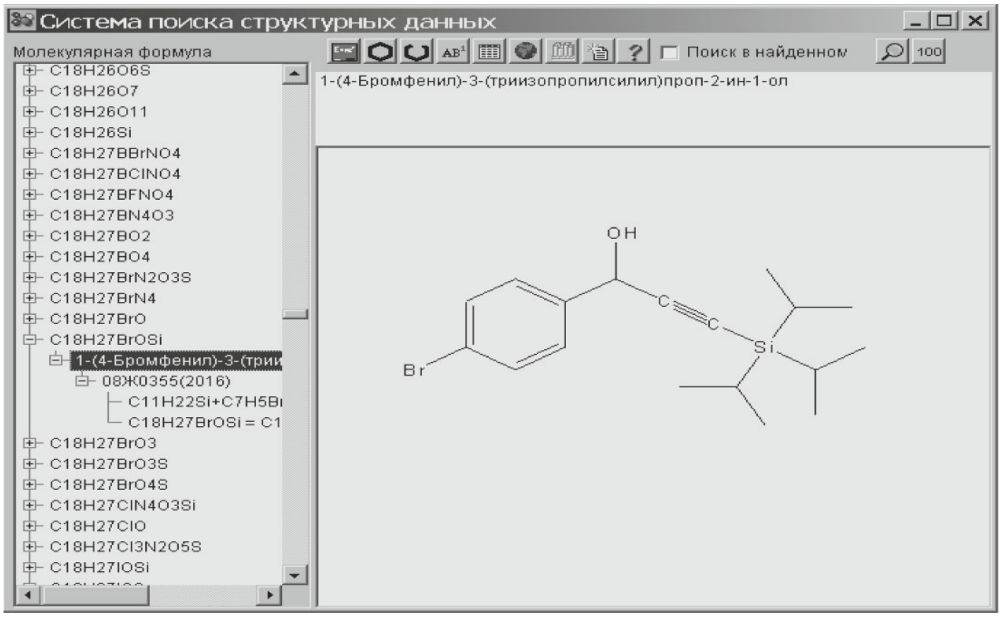
The SDB is a set of files of the same structure in a binary format (CBASE32 format), each of which contains flat tables and in addition to the data described above comprises complementary information, including service information. Files are grouped in directories that correspond to the years and numbers of the abstract database (DB) and the abstract journal (RZh) *Khimiya* (Chemistry). This organization does not allow effective searching for necessary information. The requirement for a quick information search is met by a database that is built hierarchically. To create it, a program called **Data Converter** has been developed. The main task of the converter is to create a data hierarchy focused on fast execution of queries in accordance with the technical requirements listed above. Since the hierarchical database created by **Data Converter** is designed to provide chemical information to users, it will be referred to as the user database in the future.

The **Data Converter** works as follows. Data from the SDB are normalized and acquire a four-level hierarchy. The highest level is the molecular formula, followed by the chemical structure (stereoisomers are considered different chemical structures) and its name. At the third level of the hierarchy is a list of abstracts from the *Khimiya* (Chemistry) database, in which this chemical structure is mentioned, along with a list of subject characteristics, and a bibliography. Finally, at the fourth level, each number of the abstract is assigned with a list of reactions from the article corresponding to this abstract. Molecular formulas (the first level of the hierarchy) are formed according to Hill's rule [5] and sorted alphabetically. At the second level of the hierarchy, names along with chemical structures are sorted alphabetically. Sorting is used for a list of abstracts as well. This way of presenting data allows the user to quickly find the information of interest in manual mode without creating search queries.

When creating the hierarchy, lists of unique molecular formulas are created and lists of chemical structures are formed inside each list, for which the InChIkey string [6] is used, which uniquely characterizes the chemical structure with regard to stereochemical information. Abstract numbers are used to create lists of abstracts.

The program, together with information about chemical compounds, is installed on a local computer. Data for the program are stored in a local file. This storage method allows avoiding the installation and administration of SQL server applications. An important advantage of SQL server applications, that is, multi-user capability, does not matter in this case, since the data are not edited by clients; this allows clients to access these data from several local copies of applications without causing conflicts.

The hierarchy levels listed above contain strings as a unique identifier: abstract numbers for a list of abstracts, a molecular formula for a list of formulas, and InChIKey for a list of chemical structures. When converting flat tables, it is necessary to check for the presence of this string in already formed lists. If the string is present, the data are added to the existing list, while if there is no string, a new list is created. The search time for a string in the list is proportional to $\log_2 N$ for a sorted list and N for an unsorted list, where N is the number of items in the list. Obviously, searching in sorted lists is more efficient. However, when creating a new list, the string must be included in the search; this requires re-sorting and time costs in proportion to $N\log_2 N$. If re-sorting is carried out after adding each new element, then there is no advantage of searching in a sorted list. Therefore, for the formation of lists, two arrays, sorted and unsorted, are used. The search is carried out through bisections in a sorted array (the convergence rate is $\log_2 N$) and serial searching (the rate is proportional to N) in an unsorted array. When creating a new list, an identifier string is added to an unsorted array. When the number of elements in an unsorted array reaches 64, they are transferred to a sorted array, which is re-sorted, and the unsorted array is reset. This procedure significantly reduces the time for generating hierarchical lists.



```
⊞ C18H26O6S
⊞ C18H26O7
⊞ C18H26O11
⊞ C18H26Si
⊞ C18H27BBrNO4
⊞ C18H27BClNO4
⊞ C18H27BFNO4
⊞ C18H27BN4O3
⊞ C18H27BO2
⊞ C18H27BO4
⊞ C18H27BrN2O3S
⊞ C18H27BrN4
⊞ C18H27BrO
⊟ C18H27BrOSi
    ⊟ 1-(4-Бромфенил)-3-(трии
        ⊟ 08Ж0355(2016)
            ├ C11H22Si+C7H5Bı
            └ C18H27BrOSi = C1
⊞ C18H27BrO3
⊞ C18H27BrO3S
⊞ C18H27BrO4S
⊞ C18H27ClN4O3Si
⊞ C18H27ClO
⊞ C18H27Cl3N2O5S
⊞ C18H27IOSi
```

1-(4-Бромфенил)-3-(триизопропилсилил)проп-2-ин-1-ол

Fig. 1. The main window of the Module interface of the autonomous system of structural search. The name and structure of a chemical compound are displayed.

We studied the dependence of the creation time of a final sorted list on the number of elements in an unsorted array after reaching which a single array is generated. The curve that reflects this relationship has a minimum; however, the optimal number of elements for which it is necessary to create a single array and perform re-sorting depends on the list length and the frequency of new list elements and is the subject of a separate study.

Due to a specific feature of data storage in the VINITI RAS, which consists in the fact that the bibliographic information and the abstract numbers assigned in the RZh and the *Khimiya* (Chemistry) database are stored separately from the array of structural chemical information, the converter program reads these data from external sources and inputs them in CBASE32 format tables.

As an input, the converter program receives lists of tables and generates a binary local database file for an autonomous structural search system. The **Data Converter** is implemented in 64-bit architecture, which makes it possible to convert a very large amount of data. In practice, the entire database of structural and chemical information, containing more than 8 MB of records, was successfully converted into a user database.

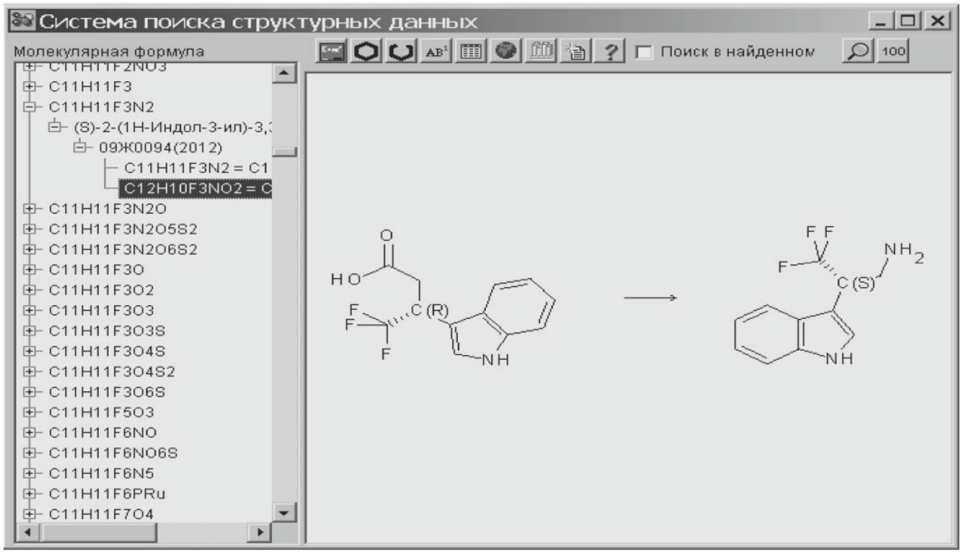
In addition, the **Data Converter** can be successfully used to create thematic databases. A thematic database is a user database that contains information about compounds with specified subject characteristics. They can be specified in the converter program before forming the base. The result is a filtered base that contains chemical structures with given terms and corresponding bibliographic data. Chemical reactions are not included in the thematic databases, as for their display, in general, all the chemical structures of this publication are necessary and some of them are filtered.

THE FEATURES OF THE SEARCH FUNCTIONS OF THE SYSTEM

The autonomous structural search system consists of two executable modules: the **Data Converter** and the **Search Module**. The latter implements the functions of searching for chemical structures in a local database formed from the SDB files using a converter program.

The main window of the **Search Module** interface of the system includes the display area of the molecular formula tree and the toolbar. The molecular formula tree reflects the four-level database hierarchy formed by the converter program, as described above.

One can browse the tree using the scroll bars or the wheel of a computer mouse. Tree nodes are expanded or collapsed with a mouse click. The tree has up to four levels. The transition to the levels below the first one is accompanied by the display of information in the right part of the window. At the first level of the tree are molecular formulas ordered according to Hill's rule. At the second level are the names of chemical compounds and their structure (Fig. 1). At the third level are bibliographic and subject information. Nodes of the third level are assigned with the abstract number in the *Khimiya* (Chemistry) database and brackets indicate the year of publication (see Fig. 11). At the fourth level (if



C11H11F2NO3
C11H11F3
C11H11F3N2
(S)-2-(1H-Индол-3-ил)-3,3
09Ж0094(2012)
C11H11F3N2 = C1
C12H10F3NO2 = C
C11H11F3N2O
C11H11F3N2O5S2
C11H11F3N2O6S2
C11H11F3O
C11H11F3O2
C11H11F3O3
C11H11F3O3S
C11H11F3O4S
C11H11F3O4S2
C11H11F3O6S
C11H11F5O3
C11H11F6NO
C11H11F6NO6S
C11H11F6N5
C11H11F6PRu
C11H11F7O4

Fig. 2. An example of displaying information about reactions.

there is one) are the equations of chemical reactions in which a given compound participates (Fig. 2).

The toolbar consists of control elements for the implementation of various types of search and execution of auxiliary actions (Fig. 3).

The search for the molecular formula and the name of the chemical compound is carried out by specifying a query in text form in the field of special windows called by the E=mc and AB¹ buttons, respectively. As an example, if one clicks the button first on the left and specifies the string C6H4 (Fig. 4) in the text box of the query window, one obtains the result shown in Fig. 5.

It can be seen that not only was the cyclohexa-1,3-dienine compound with the molecular formula C6H4 found, but also 289 chemical compounds whose molecular formula contains C6H4 as a fragment.

Similarly, the search is carried out by the systematic or trivial name of the chemical compound. In the query, one can specify the full name or its fragment in Russian.

Searching for the exact structure is carried out using the toolbar button ⬡. It calls the window of the graphic editor [7], in which one can form a structural query in the form of a chemical structure drawing. An example of such a query is presented in Fig. 6.

When the drawing is complete, the searcher clicks OK. The program will perform a search, whose results will appear in the main interface window. The search algorithm is based on the use of the InCHIKey, which identifies a chemical compound with a high degree of uniqueness.

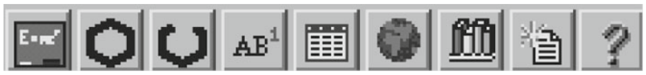


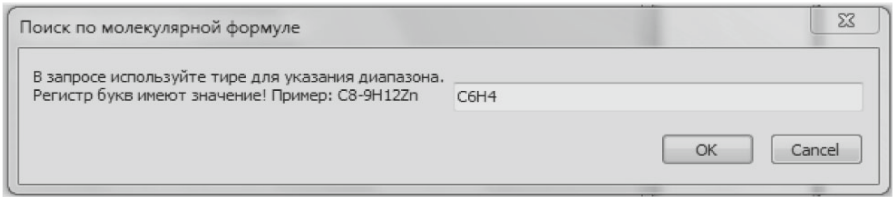
Fig. 3. The toolbar view.

The InChI string is recommended by IUPAC as a linear representation of the chemical structure [8]. It contains data on the bonds between atoms, as well as stereochemical information and is unique for identical structures regardless of the order in which the atoms are numbered (the canonical representation). For simple tautomers arising from 1,3-migrations of the hydrogen atom, as well as chains of such migrations, identical strings are generated. Thus, by comparing InChI strings, it is possible to determine whether the chemical structures are the same or different.

The disadvantage of the InChI string is that it has a variable length and can reach several thousand characters for large compounds. This flaw has been overcome by the InChIkey technology. InChIkey is an InChI hash with a length of 27 characters. The first 14 characters contain information about the bonds; after the separator stereochemical information (ten characters) and a checksum/version of InChI (one character) occur. By comparing the first 14 characters, one can find chemical structures with the same topology. By comparing ten stereochemical characters one can find identical stereoisomers.

There is a non-zero probability that two different chemical structures generate the same hash code. However, it is extremely low; at present it was not possible to reliably detect such an event for any pair of chemical structures. The matches discussed in the literature are explained by different implementations of the InChI string generation due to errors in the programs [9].

When the search program starts, the InChIkey strings are sorted and loaded into RAM. As well, for a structural user request, InChIkey is generated and the

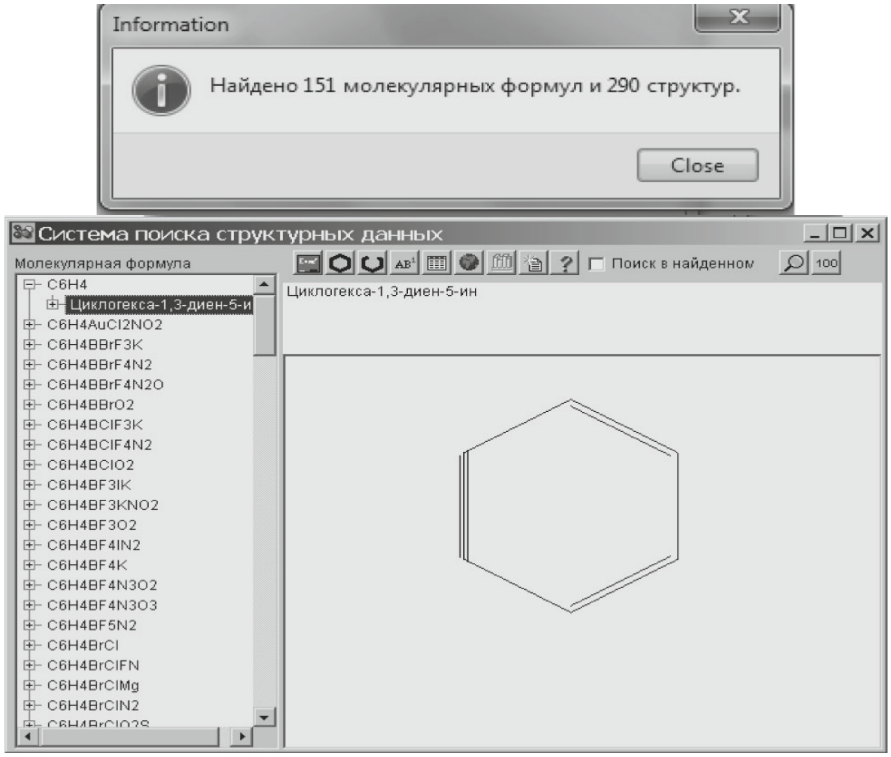


Поиск по молекулярной формуле

В запросе используйте тире для указания диапазона.
Регистр букв имеют значение! Пример: C8-9H12Zn

C6H4

Fig. 4. The window of a search query via a molecular formula.



Information

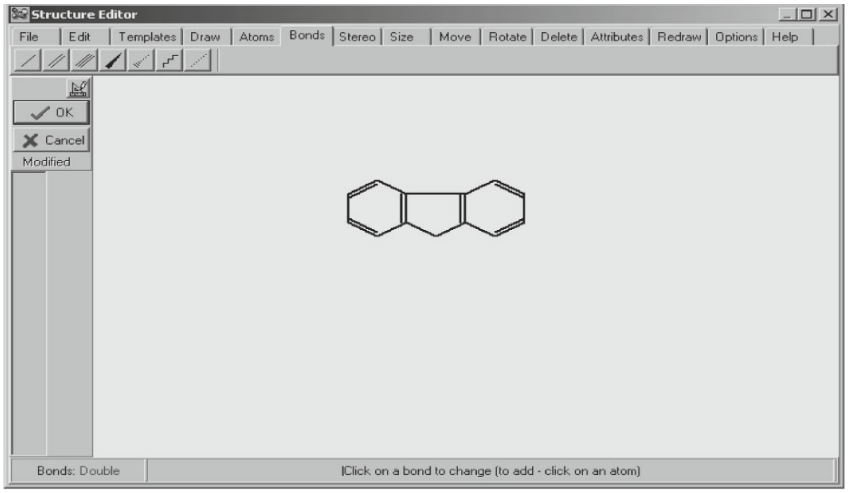
Найдено 151 молекулярных формул и 290 структур.

Молекулярная формула

```
⊟ C6H4
   ⊞ Циклогекса-1,3-диен-5-и
⊞ C6H4AuCl2NO2
⊞ C6H4BBrF3K
⊞ C6H4BBrF4N2
⊞ C6H4BBrF4N2O
⊞ C6H4BBrO2
⊞ C6H4BClF3K
⊞ C6H4BClF4N2
⊞ C6H4BClO2
⊞ C6H4BF3IK
⊞ C6H4BF3KNO2
⊞ C6H4BF3O2
⊞ C6H4BF4IN2
⊞ C6H4BF4K
⊞ C6H4BF4N3O2
⊞ C6H4BF4N3O3
⊞ C6H4BF5N2
⊞ C6H4BrCl
⊞ C6H4BrClFN
⊞ C6H4BrClMg
⊞ C6H4BrClN2
⊞ C6H4BrClO2S
```

Циклогекса-1,3-диен-5-ин

Fig. 5. An example of a result of searching via a molecular formula.



Bonds: Double

|Click on a bond to change (to add - click on an atom)

Fig. 6. The graphics editor window.

Query:
Result:

Fig. 7. An example of a query for a fragment and a structure that was found.

search is carried out in the array using a bisection algorithm. Searching through very large amount of data takes negligible time and the results are immediately presented to the user. A detailed description of the search system in the SDB on the basis of InChIkey was given in [10].

Searching for a structural fragment is carried out by the button. This type of search is of great importance in determining the relationship of the structure and activity of chemical compounds. The search algorithm is based on backtracking in combination with the relaxation method and was described in detail in [11]. It should be noted that in terms of graph theory, here we solve the problem of determining an isomorphism to a partial subgraph [12]. As an example, using the query shown in Fig. 7 the structure of the cyclohexane molecule can be found.

The search request for a structural fragment is specified using all available graphic tools of the drawing of the exact structure. In addition, there are special tools for specifying a search by fragment: *search atoms* and *search bonds*. These are represented by a group of *Query* commands of the graphical editor. The query must contain at least two non-hydrogen atoms (this is a natural and generally accepted requirement) and must not contain more than 1000 atoms and 1000 bonds in a connected fragment (this is a limitation of the autonomous structural search system). *Search atoms* are substituent atoms used to draw a query to search by substructure. The following atoms are defined as search atoms:

1. *Any*: any atom except for a hydrogen atom.

2. *Hetero*: *N*, *O*, *Si*, *P*, *S*, *Ge*, *As*, *Se*, *Sb*, *Te*, and *Po*.

3. *Exactly coordinated* is used in combination with any atom and means that the coordination of the atom cannot be different from that indicated.

4. *Halogens*: *F*, *Cl*, *Br*, *I*, and *At*.

5. *Metal*: *Li*, *Be*, *Na*, *Mg*, *Al*, *K*, *Ca*, *Sc*, *Ti*, *V*, *Cr*, *Mn*, *Fe*, *Co*, *Ni*, *Cu*, *Zn*, *Ga*, *Rb*, *Sr*, *Y*, *Zr*, *Nb*, *Mo*, *Tc*, *Ru*, *Rh*, *Pd*, *Ag*, *Cd*, *In*, *Sn*, *Cs*, *Ba*, *La*, *Ce*, *Pr*, *Nd*, *Pm*, *Sm*, *Eu*, *Gd*, *Tb*, *Dy*, *Ho*, *Er*, *Tm*, *Yb*, *Lu*, *Hf*, *Ta*, *W*, *Re*, *Os*, *Ir*, *Pt*, *Au*, *Hg*, *Tl*, *Pb*, *Bi*, *Fr*, *Ra*, *Ac*, *Th*, *Pa*, *U*, *Np*, *Pu*, *Am*, *Cm*, *Bk*, *Cf*, *Es*, *Fm*, *Md*, *No*, and *Lr*.

Search atoms of types 1, 2, 4, and 5 are used as ordinary physical atoms. Search atoms of type 3 should be drawn in combination with others (physical and search ones). *Search bonds* are substituent bonds used to represent the task of searching by a structural fragment. The editor defines the following search bonds:

1. *Any* means that two atoms are connected by an arbitrary bond.

2. *Aromatic* means that a pair of atoms is part of an aromatic ring.

3. *S/D*, Single or Double. This means that atoms are connected by a single or double bond.

4. *Ch*, Chain. This is used in combination with other bonds (physical and search ones) and means that the bond is not included in the ring.

5. *Rn*, Ring. This is used in combination with other bonds (physical and search ones) and means that the bond is part of a ring.

6. *Cis/trans*. This is used in combination with other bonds and indicates the type (*Cis/trans*) of the tagged bond.

Let us illustrate the use of search atoms and bonds based on the example of a search query, in which *Rn* denotes a cyclic bond and *X* is a halogen atom (Fig. 8). One of the results of this search is shown in Fig. 9.

Searching by subject characteristics is carried out using the button. Terms are denoted by a single character or their sequence, up to four characters long, represented in the right part of the window (Fig. 10).

The lower part of the window contains tools for searching terms by a given fragment of a term string or comment text. After the term is selected and checked, pressing the OK button will start the search procedure. An example of possible results is shown in Fig. 11.

The search for the selected structure on the Internet is carried out via the button. Clicking it forms the InChiKey character string of this chemical compound, which is automatically placed in the Google search box of the Internet browser.

To view data from the electronic catalog of the VINITI RAS, it is necessary to select a tree node with the abstract number in the *Khimiya* (Chemistry) database (Fig. 12).

The button then becomes active; upon clicking it, a connection with the Electronic Catalog of the VINITI RAS is established. The Internet browser displays the bibliographic information of the original source and provides access to the Electronic Catalog [13], where one can obtain additional information:

- translation of the name of the source into Russian;
- date of registration in the VINITI;
- storage place in the VINITI;
- annotation;
- keywords;
- references to all publications of each of the authors of a given article.

The Electronic Catalog Service on the Search tab allows one to find bibliographic information by keywords and other parameters.

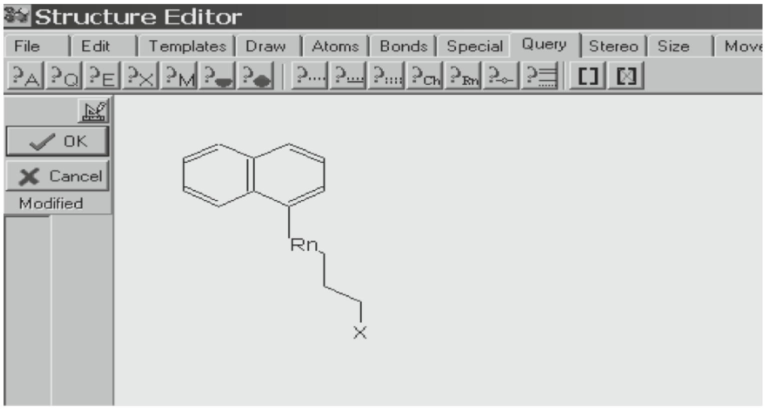
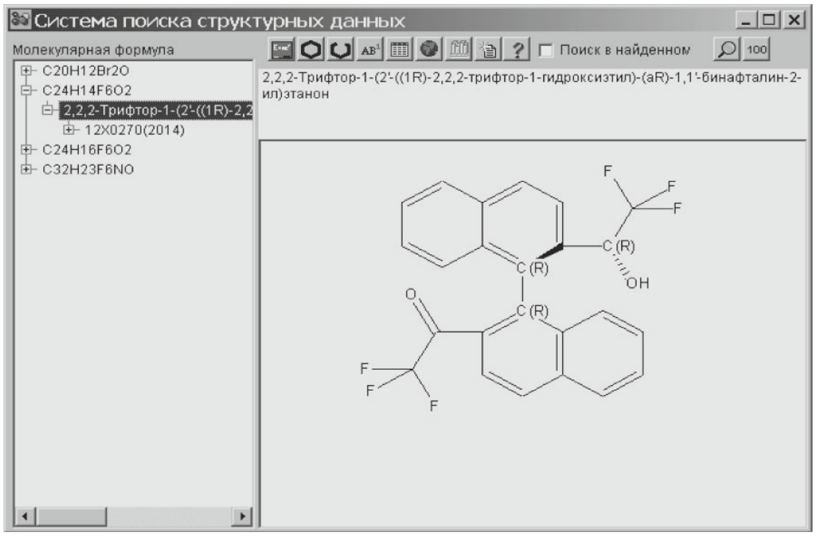


Fig. 8. An example of a search query via a structural fragment.



- C20H12Br2O
- C24H14F6O2
  - 2,2,2-Трифтор-1-(2'-((1R)-2,2
    - 12X0270(2014)
- C24H16F6O2
- C32H23F6NO

2,2,2-Трифтор-1-(2'-((1R)-2,2,2-трифтор-1-гидроксиэтил)-(aR)-1,1'-бинафталин-2-ил)этанон

Fig. 9. One of the results of searching via the fragment in Fig. 8.

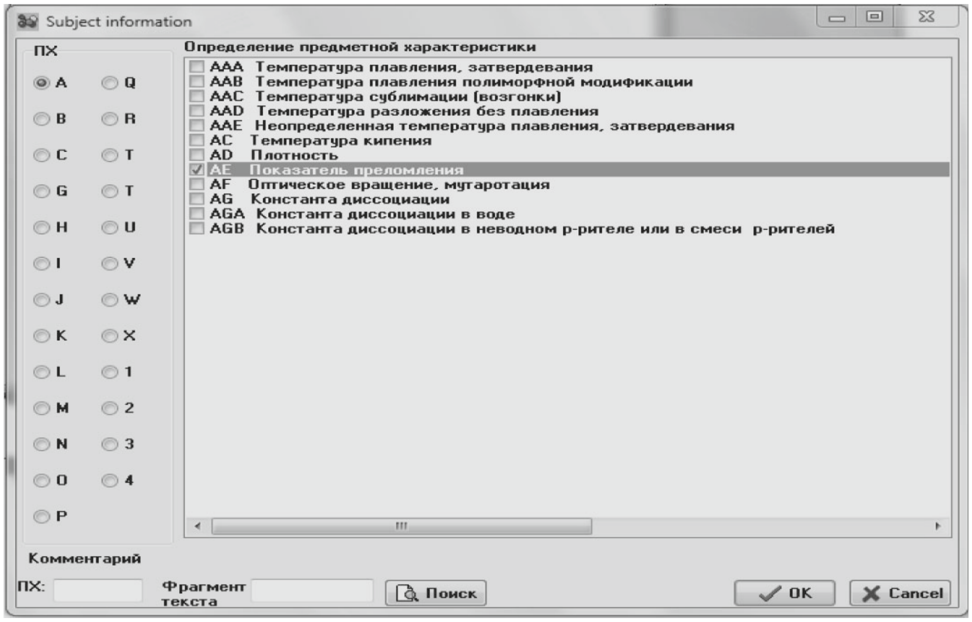
CONCLUSIONS

Currently, the problem of providing users with information from the structural chemical information database is solved by passing them to a user database and the **Search Module** program described above. The user database is created using a special data converter program, where the data are systematized by the periods during which the information was collected. As an example, the freely distributed demo version of the autonomous search system [14] contains an array of data from the SDB (approximately 200000 chemical structures) accumulated over 2012.

Further development of the system can follow in two directions.

The first direction is an increase in the amount of information provided to the user due to retro data. It should be noted that the task of converting data for the 1996–2009 period (27% of the total number of chemical compounds) has been solved; however, the conversion of data from 1975–1995 (62% of the total number of chemical compounds) is a rather time-consuming task. The problem is that compounds in this period were input without using a graphical editor; thus, the database files do not store information about the structural formula drawing. The task of automatically constructing a drawing of the structure according to the table of bonds (visualization of the molecule) is currently successfully solved for most of the chemical compounds of the SDB [3].

The second direction involves the provision of thematic samples from the SDB (for example, biologically active compounds, complex compounds, compounds used in a particular technology, etc.) at the request of users. Currently, the task of creating specialized local databases in the autonomous search system has been partially solved on the basis of the SDB: a converter has been developed to form thematic data-



Subject information

ПХ

● A ○ Q ○ B ○ R ○ C ○ T ○ G ○ T ○ H ○ U ○ I ○ V ○ J ○ W ○ K ○ X ○ L ○ 1 ○ M ○ 2 ○ N ○ 3 ○ O ○ 4 ○ P

Определение предметной характеристики

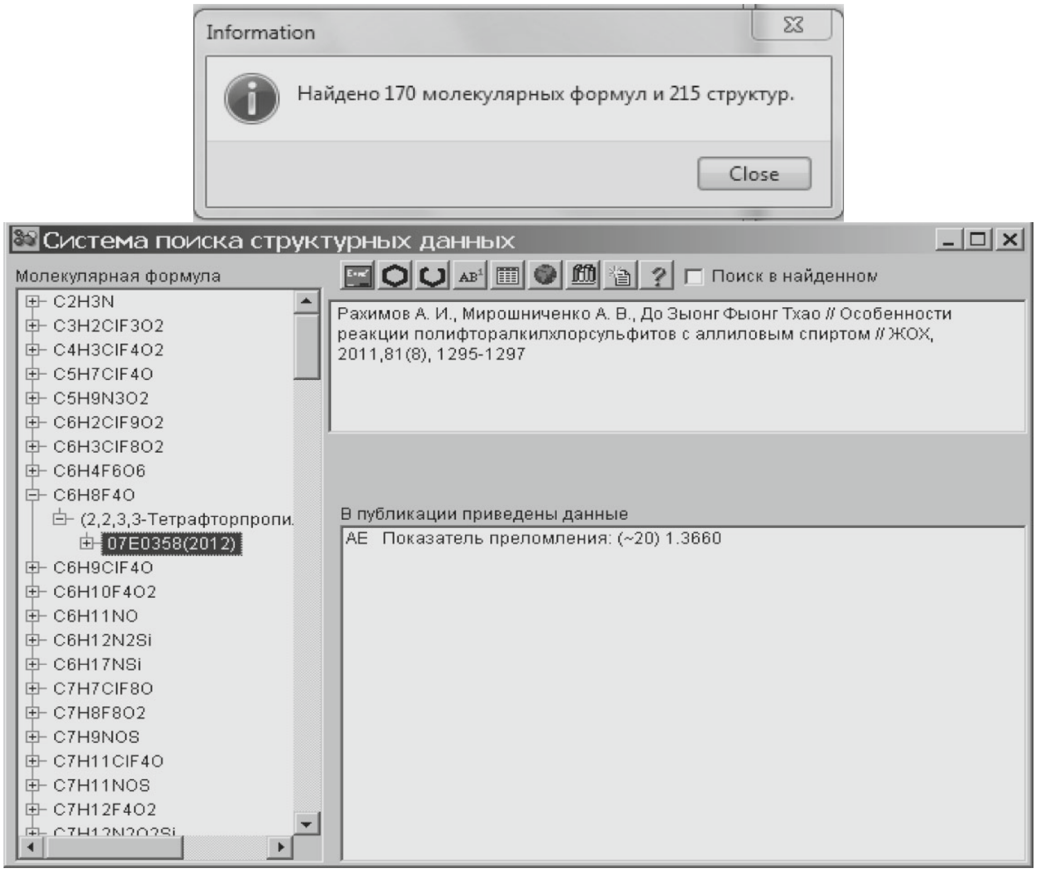
- □ AAA Температура плавления, затвердевания
- □ AAB Температура плавления полиморфной модификации
- □ AAC Температура сублимации (возгонки)
- □ AAD Температура разложения без плавления
- □ AAE Неопределенная температура плавления, затвердевания
- □ AC Температура кипения
- □ AD Плотность
- ☑ AE Показатель преломления
- □ AF Оптическое вращение, мутаротация
- □ AG Константа диссоциации
- □ AGA Константа диссоциации в воде
- □ AGB Константа диссоциации в неводном р-рителе или в смеси р-рителей

Комментарий

ПХ:

Фрагмент текста

Fig. 10. The window for selecting a characteristic of a subject.



Information

Найдено 170 молекулярных формул и 215 структур.

Система поиска структурных данных

Молекулярная формула

- C2H3N
- C3H2ClF3O2
- C4H3ClF4O2
- C5H7ClF4O
- C5H9N3O2
- C6H2ClF9O2
- C6H3ClF8O2
- C6H4F6O6
- C6H8F4O
  - (2,2,3,3-Тетрафторпропи
    - 07E0358(2012)
- C6H9ClF4O
- C6H10F4O2
- C6H11NO
- C6H12N2Si
- C6H17NSi
- C7H7ClF8O
- C7H8F8O2
- C7H9NOS
- C7H11ClF4O
- C7H11NOS
- C7H12F4O2
- C7H12N2O2Si

□ Поиск в найденном

Рахимов А. И., Мирошниченко А. В., До Зыонг Фыонг Тхао // Особенности реакции полифторалкилхлорсульфитов с аллиловым спиртом // ЖОХ, 2011,81(8), 1295-1297

В публикации приведены данные

AE Показатель преломления: (~20) 1.3660

Fig. 11. An example of a result of searching via a subject characteristic.

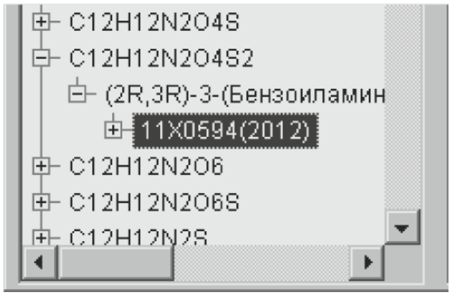


Fig. 12. A selected node with an abstract number.

bases that contain information about compounds with specified subject characteristics.

The considered system of autonomous information search in user databases can be applied in the following areas:

• fundamental and applied research in chemistry, biochemistry, pharmaceuticals, pharmacology, toxicology, ecology, and medicine;

• the chemical industry;

• education (assistance to teachers, students, graduate students, and professionals who are raising their qualifications);

• information support (databases, scientific libraries, and publishing houses).

REFERENCES

1. Some scientometric data in chemistry. http://chem21.info/article/388814/.
2. Alfimov, M.V., Avakyan, V.G., Trepalin, S.V., Voronezheva, N.I., and Churakova, N.I., A universal software shell for creating databases of chemical compounds and reactions, *Dokl. Ross. Akad. Nauk.,* 1999, vol. 366, no. 5, pp. 639–642.
3. *Baza strukturnykh dannykh po khimii VINITI. Indeksirovanie i vvod svedenii o khimicheskikh soedineniyakh pri podgotovke bazy strukturnykh dannykh po khimii s ispol'zovaniem programmnogo kompleksa CBASE32. Instruktsiya VINITI RAN I81-2010* (VINITI Database of Chemical Structural Data. Indexing and Input of Information about Chemical Compounds in Preparing the Database of Chemical Structural Data Using the CBASE32 Software Package. VINITI RAS Instruction I81-2010), Moscow, 2010.
4. Nefedov, O.M., Koroleva, L.M., Trepalin, S.V., Bessonov, Yu.E., and Churakova, N.I., The development of an integrated system for structural chemical information, *Autom. Doc. Math. Linguist.,* 2015, vol. 49, no. 6, pp. 213–220.
5. Hill, E.A., On a system of indexing chemical literature; Adopted by the Classification Division of the U.S. Patent Office, *J. Am. Chem. Soc.,* 1900, vol. 22, no. 8, pp. 478–494.
6. Heller, S., McNaught, A., Stein, S., Tchekhovskoi, D., and Pletnev, I., InChI—the worldwide chemical structure identifier standard, *J. Cheminf.,* 2013, vol. 5, no. 1, p. 7.
7. Description of the Graphical Editor CBASE32. https://yadi.sk/i/qMWnyrNFzjkH5.
8. The IUPAC international chemical identifier. https://iupac.org/who-we-are/divisions/division-details/inchi/.
9. An InChIkey Collision is Discovered and NOT Based on Stereochemistry. http://www.chemconnector.com/2011/09/01/an-inchikey-collision-is-discovered-and-not-based-on-stereochemistry/.
10. Nefedov, O.M., Trepalin, S.V., Koroleva, L.M., and Bessonov, Yu.E., Fast search for exact chemical structures in large databases using InChI key coding of structures, *Nauchno-Tekh. Inf., Ser. 2,* 2013, no. 12, pp. 27–33.
11. Nefedov, O.M., Trepalin, S.V., Koroleva, L.M., Bessonov, Yu.E., and Churakova, N.I., The database of chemical structural data of the VINITI RAS: Problems of searching for structural fragments, *Nauchno-Tekh. Inf., Ser. 2,* 2014 no. 12, pp. 19–29.
12. Berge, C., *Théorie des Graphes et Ses Applications,* Dunod Orléans, 1963, 2nd ed.
13. VINITI RAS Electronic Catalog. http://catalog.viniti.ru/.
14. Demo Version of Autonomous Structural Search in the Database of Chemical Compounds. https://yadi.sk/d/JWUaGwPMvb6Mq.

Translated by K. Lazarev