CheD: Chemical Database Compilation Tool, Internet Server, and Client for SQL Servers

Sergei V. Trepalin† and Alexander V. Yarkov*,‡

Institute of Physiologically Active Compounds, 142432 Chernogolovka, Moscow Region, Russia

Received April 30, 2000

An efficient program, which runs on a personal computer, for the storage, retrieval, and processing of chemical information, is presented. The program can work both as a stand-alone application or in conjunction with a specifically written Web server application or with some standard SQL servers, e.g., Oracle, Interbase, and MS SQL. New types of data fields are introduced, e.g., arrays for spectral information storage, HTML and database links, and user-defined functions. CheD has an open architecture; thus, custom data types, controls, and services may be added. A WWW server application for chemical data retrieval features an easy and user-friendly installation on Windows NT or 95 platforms.

1. INTRODUCTION

In recent years database development has progressed in two directions. First, SQL technology[1] is widely used for the development of databases. SQL servers give unique possibilities such as integration of data with other databases, multiuser access for data modification, and easy writing and modification of client applications. But, databases created with SQL technology are difficult to distribute between different locations. A distribution package has to include, at least, client routines for the selected SQL server. The user should have sufficient qualifications to install and administer the client routines for most SQL servers. Besides, most companies, the owners of SQL servers, do not give a license for the distribution of client routines to third parties. As a result, SQL databases can be treated only as corporate databases, which are located at a restricted number of places. Second is the tendency for database exposure via the Internet.[2] The Internet is a relatively new technology, and it is only recently that standard routines for Internet data exposure/searching have been included in packages for some SQL servers. Development of the server application is a relatively simple procedure for data which can be presented to the client as a text. But, the presentation and retrieval of chemical information, e.g., structures and spectral data, require much more effort than usual data processing.

Despite these general trends, local databases still play a significant role in the presentation of chemical information—for example, the distribution of data on CD-ROM. SQL data cannot be distributed as such, because a SQL server has to be installed at a desirable location (at least, the client part of the SQL server). In addition, staff with appropriate knowledge are required to provide support for the SQL server. At the moment we are not aware of any publicly available Web site, which provides the ability to upload and modify the content of a chemical information database. This means that individual chemists and small companies need local databases and applications to compile them.

† E-mail: trep@ism.ac.ru.
‡ E-mail: yarkov@ipac.ac.ru. Phone: 7-095-5245062.

It is obvious that the availability of easy-to-use tools for data conversion among these three types of chemical data storage, i.e., local, corporate, and WWW databases, would minimize errors and reduce the cost of data processing. This problem has already been solved in some modern products. For example, ISIS/HOST (Molecular Design Ltd.)[3] has connections to the Oracle SQL server. Chemscape server and Chime Pro Suite allow a Web server to communicate with ISIS/HOST, enabling developers to provide chemical structure rendering and visualization in pure Java applications and applets as well as Web pages. ChemOffice Web server[4] connects to Microsoft and Oracle SQL servers and supports local databases. The data collected with ChemOffice might be directly used by the ChemFinder WebServer.[5]

The CheD system partially solves the problem of data conversion among all three kinds of data storage too. It works as a stand-alone application and as an Internet client. ISAPI DLL has been written for use with CheD. The file ChemWSer.dll is installed on a computer running the Microsoft Web server. It transfers data to the CheD program running on a remote computer and generates both HTML pages for any Web browser and HTML pages with references to ActiveX controls for use with Microsoft Internet Explorer. The service ChemSQL.dll is used to import/export/browse/edit chemical data stored on Oracle, Interbase, or MS SQL servers. The data from a SQL database can be easily integrated into a corporate database using the tools supplied with the CheD software.

Besides this novelty, CheD has another feature, also present in specific programs such as ACDLabs Spec Manager.[6] This is spectral database management, prediction, and Internet exposition.

2. DATABASES

A database size is limited to 10 GB, with the maximum number of records being 10 000 000. Each database can contain up to 500 data fields. Their types and descriptions are shown in Table 1. Data associated with chemical structure are stored in such a way that *chemical structure* is a

10.1021/ci000039n CCC: $20.00 © 2001 American Chemical Society
Published on Web 12/14/2000

Table 1. Field Types Supported by CheD[a]

N	type	maximal array dimensions	range flag available?	is field searchable?	assigned to atoms?	available in ISIS?[3]
1	chemical structure	1	−	+	−	+
2	Brutto formula	0	−	+	−	+
3	real (floating point)	9	+	+	+	+
4	real + units of measure	9	+	+	+	−
5	integer	9	+	+	+	+
6	Boolean	9	−	+	+	+
7	string (single text line)	9	+	+	+	+
8	text (multiply lines)	0	−	+	−	+
9	list	9	−	+	+	−
10	date	9	+	+	+	+
11	picture (Bitmap or Metafile)	9	−	−	−	−
12	OLE object	0	−	−	−	+
13	packed array	0	−	+	−	−
14	calculated (real value returns)	0	−	+	−	+
15	user-defined function	0	−	−	−	−
16	WWW link	1	−	+	−	−
17	database link	1	−	+	−	−
18	record	1	−	−	−	+
19	custom data	0	−	+\−	−	+

[a] Array dimensions: 0, scalar data; 1, vector; greater than 1, matrix with the maximal number of columns. "Assigned to atoms" means that values may be associated with some atoms in the chemical structure.

mandatory field. Counterions are presented as a part of the structure. Manipulations by way of chemical reactions, however, are not supported. Standard search methods are used for the usual data fields. For example, one can execute an exact structure search, a substructure search, or a similarity search of data stored in the *structure* field. Brief descriptions of the data types unique to CheD are shown below.

2.1. Database Fields. *Real + Units of Measure.* This data type enables the recalculation of data during the run time from one unit of measure to another, for example, heat capacity from kcal/mol to kJ/mol. Units of measure are defined in lists, the first element of which is assumed to be the main unit. Other list elements have linear recalculation parameters *A* and *B*. All field values, except for those measured in the main measurement unit, are recalculated using the equation

$$\text{value (unit of measure)} = A + B[\text{value (main unit of measure)}]$$

In the example above (heat capacity), the coefficient *A* is always zero. Recalculation with a shift (*A* is not zero) is required when, for example, chemical shifts in NMR spectra are recalculated to another standard. Search of these data requires a definition of the unit of measure. If the data in some records have another unit of measure, they are recalculated to match the search measure.

Lists. Some kinds of information in a database, e.g., solvents, journal titles, etc., can be represented with different strings, which have identical meanings. For example, one may define the solvent acetone as ACETONE or (CH3)-2CO. Such representations make the data difficult to search. To eliminate this problem, the information is presented as string data in indexed lists with an index key associated with each string. This has the advantage of both conserving disk space and ensuring the standardization of spelling and thus completeness in searches. The usual drawback of this method is the incompatibility of data entered at different work places. To overcome this incompatibility, CheD has a routine which compares data from two sources and creates a common list and database. A list of predefined strings is provided to the user when building the query.

User-Defined Function. This data field is similar to the usual *calculated* field and gives the user the ability to define his own equation for the value calculation. But, contrary to a *calculated* field, a variable is defined for a *function* field. The range is specified for the variable. The result of the calculation is presented as a two-dimensional plot on the screen.

WWW Link. This field is a combination of two strings: displayed text and the associated URL address. When a user clicks the text, the content of the associated URL is retrieved. When a database is exposed over the WWW (see below), this type of field is shown as a hyperlink in an HTML document. A search is performed as per regular text fields: one can define a fragment of a string in the address or in the displayed text.

Database Link. This field defines the reference to a record in another CheD database. It consists of a combination of three fields: (1) displayed text; (2) the name and path of a database; (3) reference to a record. The reference may be to a chemical structure, an integer, a string field value, or a record number. Clicking on a text will cause the appropriate database to be opened and the referenced record to be displayed. Searching on the data is performed as for regular text fields.

Packed Array. This field is manipulated with (*X*, *Y*) pairs of values (e.g., a spectrum), when the *X* values are equidistant. The *packed array* field stores *Y* values, and a starting and ending *X* value. The memory requirements are reduced by a factor of 2. The Spline approximation (see below) is used for searching this data type.

Record. The *record* does not contain any data by itself, but rather connects some data fields into a single field. It is an attribute for a group of fields. The fields from Table 1, which have an *array size* value greater than zero, can be inserted into a *record*. A *record* can have an array flag. The addition of a new element into such a record means the

addition of a single element into all the fields defined in this *record.*

Custom Data. CheD can support virtually any kind of data by adding the appropriate DLL. The DLL has to have methods which either pass a metafile picture to CheD or create a child custom control in the specified CheD area. Fourier interferogram data (FID), which are registered by NMR spectrometers, are examples of these *custom* data.

Custom Search. Custom search procedures can be created in DLLs. Custom search methods can be written for the data listed in Table 1. A DLL is responsible for the dialog generation−query builder. A "mass-spectra neutral loss" search could be mentioned as an example.

Range Flag. The data with this flag are treated as a range, for example, boiling point 135−137. Inside range and outside range searches are available.

Array Size. Some chemical data can be presented as arrays. For example, an infrared spectrum can be treated as an array of pairs (wavenumber, optical density). Two values are required to characterize each single point of an IR spectrum. Results of electron density calculations can be treated as an array of *X*, *Y*, *Z* coordinates and the corresponding value. A single point can be characterized by four values in this case. CheD supports up to nine values per point. The number of values is defined in the *array size* field. All spectral data—MS, IR, NMR, ESR, etc.—can be presented as numerical data with a nonzero *array size*. CheD has a number of query builders to retrieve information for numerical data with a nonzero value of the *array size*, and, hence, executes different types of data searching. These search types include the following.

(1) The search is executed within or outside a given range. Several ranges may be entered. Each element of the array is compared to determine if it is located within the given range(s). This type of search is convenient when a data field contains a list of values, for example, boiling points for a compound, obtained from different sources.

(2) Values for each point are defined for the number of points being determined. Some components of a point could be undefined. A similarity search is executed. This kind of search is useful, for example, for mass spectra, for a peak table generated from an IR spectrum, etc.

(3) For two-point arrays (IR, NMR, ESR spectra, etc.) with nonequal *X* steps, *Y* values are approximated by executing Spline smoothing. A similarity search is used. This kind of search is useful for spectra with different resolutions or for subspectrum searches.

Hamilton's *R* factor,[7] which can vary from 0% to 100%, is used to determine query and data matching during a similarity search. After the completion of the similarity search, the retrieved records can be arranged according to *R* factor value.

Data Assigned to Atoms. The types of data listed in Table 1 can be associated with certain atoms in a chemical structure. Several atoms can be assigned a single value, the number being determined by the physical meaning of the field. For example, the chemical shift in an NMR spectrum has to be assigned to a single atom, bond length to a pair, and so on. A special data retrieval procedure was created for the numerical data assigned to atoms. A query substructure is defined, and the atoms of interest are marked. The search returns a histogram of the matching values, the mean, and the standard deviation. In addition, it allows the user to browse through the relevant records.

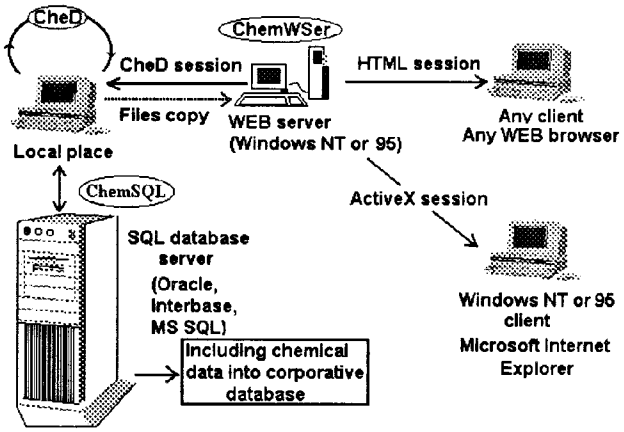

Figure 1. Architecture of the CheD program system.

2.2. Import and Export of Data. A database can be exported into another CheD database or SDF[8] or JCAMP-link[9] files. Import from these data files is also possible. Export is preceded by a fields to be exported selection dialog. Import into the existing database is started with field matching. Duplicate chemical structures are detected and displayed during import. Afterward they can be appended or skipped. JCAMP files can store IR, NMR, and MS spectral data and NMR peak table assignments. NMR peak table assignments are redefined automatically if data from a JCAMP file record are merged with an existing record in the database.

A custom-written DLL can be used for the import and export of the database records. The predefined parameters for DLL's method are the filename and list of database fields with auxiliary data—field type and array size. The DLL either executes a dialog to match fields or performs matching automatically.

The following formats are available for import and export of data for selected fields: chemical structure, MDL Molfile, JCAMP-CS,[10] Standard Molecular Data (SMD);[11] IR spectra, JCAMP, Perkin-Elmer Instrumental; mass spectra, JCAMP, Finnigan Instrumental, Nicolett Instrumental; NMR spectra, JCAMP, Aspect 2000/3000 FID (Bruker). A simple format for array data was created. Data are written into an ASCII file. The number of columns per line is equal to the array dimension (two for spectral data), and the number of lines is equal to the number of data points.

Custom formats can be added by creating an appropriate DLL. The DLL gives the format name and its type (import or export or both) and defines the data types (integer, real, string, etc.) and the array size. When a control, which contains the described data, is activated, the custom format becomes enabled in a popup menu.

3. PROGRAM DESCRIPTION

3.1. System Architecture. CheD can be considered as a suite of three applications. CheD.exe is installed on the client site. ChemWSer ISAPI DLL resides on the Web server. ChemSQL DLL is installed on the client, and the client routines for the selected SQL server are also installed on the client workstation. The program system configuration is shown in Figure 1. One can browse and edit data at the local

site in a typical way for local applications. CheD is an Internet client, and the TCP/IP protocol is used for remote access of the database located on a Web server. HTTP.ocx ActiveX control (Net Manage Inc.) is used to support the client's capabilities of CheD. The encoded binary data (not HTML documents!) are transferred during a session of CheD with the Web server. The remote databases are read-only and cannot be edited, but all other manipulations, such as search, selection, browsing, and export, are available for remote databases. Data, created at the local site, can be moved to the Web server for WWW exposure by simple copying. If ChemSQL and the client application of a SQL server are installed at the local site, data can be moved to a SQL server. In addition, if a SQL server contains any data, they can be downloaded and saved into the local database. Also, editing of the chemical data located on the SQL server is available.

ChemWSer, located on a Web server, is also used for data exposure over the WWW. Besides a session with CheD, two other types of sessions with Web clients are available. We shall call them "HTML" and "ActiveX" sessions. During these types of sessions, clients use the standard Internet browsers for the data presentation.

3.2. CheD. The stand-alone application supports chemical databases. It was created as a multiple-document interface (MDI) application. All standard commands—database field definitions and their conversion, browsing and editing of data, searching, sorting, creation of data sets and manipulations with them, import and export, clipboard operations, printing, creation of forms and tables for data presentation and printing—are available in CheD. They will not be discussed in this paper, except in the case that a command has been improved. A brief discussion of the new capabilities of CheD is presented below.

Database Header. CheD can be used to distribute chemical data on a CD-ROM. Copyrights to data depend on the owner. To manage databases from different sources, each database has a header, which contains the copyright information, contact person, E-mail, and home page location. The information is displayed on a toolbar. Clicking on E-mail or Home Page URL will launch the E-mail program or Internet browser. Additionally, the header contains information about the database's passwords. Three passwords can be defined for each database: the open password, the edit password, and the export password. The open password protects the database from unauthorized browsing, the edit password protects it from content editing, and the export password protects the database from exporting data into ASCII formatted files, e.g., SDF, JCAMP, or lists with values.

Improved Graphic Interface. A WYSIWYG graphic user interface for data presentation is used in most commercial software. This type of interface means that the data presentation on the screen is identical to their appearance on paper after printing. However, WYSIWYG technology has some disadvantages with respect to the data presentation on screen. A sheet of paper, for example, has no scroll bars, while controls on screen do. Tab controls also can be used for improvement of the data presentation. Their usage is important for displaying spectral information, because the presentation of spectral data requires a large working area. CheD has two different kinds of forms and tables, for data display and for printing. The user can edit them and define new forms or tables. Many forms and tables can be defined for a single database. Switching between them is performed by selection from a list.

Both tables and forms can be displayed on the screen at the same time. Divider movement during run time can vary their relative sizes. The typical graphic interface is shown in Figure 2. Each form has multiple pages; switching is possible by clicking on the Tab control. If a page contains a single control, it may be resized to occupy the whole client space of the page. If the size of a control is greater than the displayed area, a scroll bar will appear. Note that a single data field may have several controls on a single form. This becomes an important issue when the data might have different representations for the user. For example, a mass spectrum may be presented both as a table of m/z—intensity values and as a picture in a single form. The modification of data in one control will automatically refresh the content in the other controls defined for that data field.

Data which have a nonzero *array size* flag are shown as a table by default. For numerical data, stored as an array, special controls were created. These controls are (1) a histogram (one-dimensional array) of X, Y values as peaks (usually used to display mass spectra), (2) X, Y values with Spline approximation between points (used for display of IR, NMR, ESR, UV, and Raman spectra), (3) X, Y points with a least-squares line, and (4) spectrum simulation from intensity, half-width, and X values (three-dimensional array). In addition, user-defined controls can be included in CheD to display multidimensional data. These are discussed below.

Bookmarks. The results of a search and a database comparison are displayed as marked records. Instead of the search, one can manually put and remove a mark for any record. The content of the database can be browsed from one bookmark to another. Bookmarks can be inverted. A set can be made from the bookmarked records.

Multithreading. Commands which do not change the content of the database—searching, sorting, printing, export of data, and selection of data with lists—are executed in different threads. This gives a high level of flexibility to the program. A user, for example, can execute a search and printout of a database while at the same time editing the content of another database.

Accurate Data Compilation. To eliminate errors introduced during the data keypunching, input of the same data by two operators is effective. The problem is how to compare these two data sets for differences and, hence, possible errors. CheD has routines which compare two data sets to find differences. These data sets have to have the same number of records. They must be sorted in order that the first record of the first data set should correspond to the first record of the second data set and so on. Fields to be compared are selected by the user. Any type of field, including binary, can be compared. For binary fields (*arrays*, *OLE objects*, *pictures*) comparison is made by the bit-to-bit method. The results of a data comparison are presented as a text report and can be stored in the database(s) in a *Boolean* field.

Enhanced Data Editing. CheD can treat the content of a database as a single document. The command *Search and Replace* scans the content of the database and replaces the target value by the query value. Field types such as *string*, *text*, *integer*, *real*, *Boolean*, and *list* can be edited after responding to the requested confirmation.

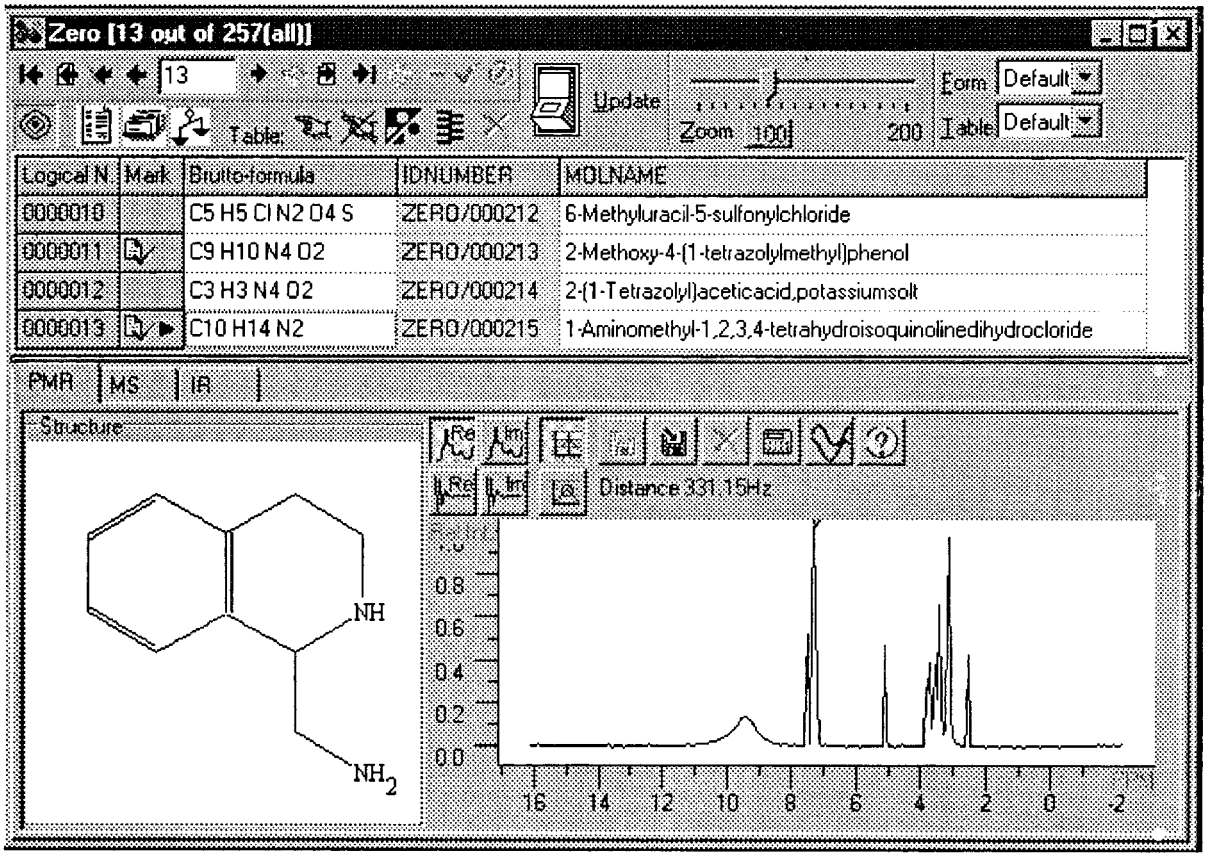

Figure 2. CheD graphical interface.

Deleting of multiple records by a single command is possible. CheD can copy both single and multiple records from one database to another with field matching.

Atom-Assigned Property Prediction. Some types of data, which are centered on atoms, have transferability from structure to structure. This means that nearly equal numerical values of a property are observed in different, but similar structures. For example, chemical shifts in NMR spectra are transferable. CheD has built-in routines for the estimation of transferable properties. In the first step, the databases having previously defined values are scanned to extract the value and structure fragments. The structural fragments are centered at the atom to which the property is assigned. If two (or more) identical fragments are found in different structures, the average value and the dispersion of that property are calculated. The fragments and the corresponding average values and the dispersions are stored in a separate file. To predict a value, the target structure is divided into fragments and the selected file is scanned for those fragments. If a fragment is found, the average value is displayed as the predicted one.

Database Comparison. Sometimes it is necessary to be aware of duplicate records in a database(s). This problem appears, for example, when external data are added to an existing database. The content of two CheD databases can be compared by Drag-and-Drop of the source database into the target. The pair of fields, which are compared, must be defined. The selected fields should be one of the following types: *chemical structure*, *Brutto formula*, *integer*, *real*, *real + units of measure*, *string*, *list*, *Boolean*, *date*. The results of the comparison are marked records, which can be analyzed and/or deleted, a text report and a table displaying record pairs with identical contents for the selected field. A database can be compared for duplicate records within itself.

Comparison by chemical structure is very fast, taking about 2 min to compare the contents of two databases with a size of 100 000 records each (P-II, 450 MHz, 128 MB of RAM). Stereo-labeled atoms and bonds are not taken into consideration during chemical structure comparison.

Access to Remote Databases. CheD is a client application which supports network access to databases. The TCP/IP protocol is used for data exchange. Because this protocol is part of the Windows operating system, no additional network configuration is required for CheD to run. To expose a database for remote access, ChemWSer ISAPI DLL, which is described below, must be installed on the server. The remote database is read-only and cannot be edited. User accessible options are search, extraction loading, and sorting. These methods are executed on the server side to reduce network traffic. A remote database can be downloaded to the local site either in internal CheD format or as SDF or JCAMP files.

Integration with Custom Software. CheD is an OLE automation server. This means that applications created with any high-level programming language can call predefined CheD commands and get and put data. Some of the commands which are accepted by CheD are Open Database, Change Current Record, Set Bookmarks, Make Set, Execute Search, Edit Current Record, and many others.

Methods defined in the third-party DLLs can be called from CheD. A DLL which can be used by CheD must expose methods with predefined names. These methods with their parameter lists are documented. To use any DLL with CheD,

it is necessary to put it into the directory where the CheD.exe file is located. CheD performs the registration of a DLL automatically. Except for the above-mentioned DLLs for custom data type definition, custom graphics, custom formats (multiple- and single-record), and custom searching, CheD supports *custom service* DLLs. These are discussed below.

Not only can CheD call methods from a DLL, but the DLL can also call methods defined in CheD by the OLE automation protocol. For example, a DLL can execute the File Open dialog, pass the filename to CheD to open it, read data from the first record, jump to the next and read the new data, etc.

3.3. Services. Services are not a part of CheD itself, but are created in the DLLs. CheD automatically registers the DLL, when it is copied to the CheD.exe directory, and the command for this service execution becomes available from the main menu of CheD.

Chemical Calculator. This service contains simple methods which are used by chemists in their routine work.

(1) Reaction calculation: Calculates from Brutto formulas (or structures) and stoichiometric equations weights and volumes of reagents and products.

(2) Calculates possible Brutto formulas from the composition and calculates an error for the composition. Valence checking is also possible.

(3) Recalculates concentration units: molar, normal, grams per liter, percentage.

(4) Solution dilution. Calculates volume, weight, and concentration for the mixture of two solutions.

(5) Calculates total, σ, and π charges for a defined chemical structure.

Chemical Structure File Processing. (1) Browsing and editing of a chemical structure in files of the following formats: MDL Molfile, MDL SDF, JCAMP-CS, SMD, HIN. For the multistructure formats (SDF, JCAMP) viewing and edition of a separate record is possible.

(2) Browsing, retrieving, and editing of the chemical structure and auxiliary data directly in an SDF file. Searches by structural fragment and by text are available. A new data field can be added; existing ones can be renamed or deleted. The Find and Replace command for a substring in a data field(s) is available.

(3) Batch converter of SCF files (ChemWindow)[12] to MDL Molfiles.

(4) Batch converter of MDL Molfiles to an SDF file.

(5) Structural similarity and diversity[13] analysis. Calculates the similarity of structures and similarity of the database to a structure. Calculates and sorts the database by diversity. Selects the most diverse proposed structures.

Mass spectrum processing is a collection of procedures and algorithms used in mass-spectral research.

(1) Calculation of masses and isotopic distributions for a given Brutto formula.

(2) Calculates the intensity of a peak for a given Brutto formula with isotopic composition. Example: $^{12}C_2{}^{13}C_1{}^{35}Cl_2{}^{37}Cl_2{}^{1}H_4$. Natural abundance is assumed.

(3) Periodic table of the elements with abundance of isotopes and their exact masses.

(4) Calculations of a Brutto formula from exact mass measurements. List of elements and maximum and minimum number of atoms that can be imposed as constraints. The best candidates are presented in a sorted mode.

(5) Calculates mass and isotopic distribution of a structural fragment defined by "lasso" or bond splitting.

IR Spectrum Processing. (1) Spectrum subtraction, absorbance to transmittance interconversion, baseline correction, and spectrum shift and normalization.

(2) Spectrum editing point by point.

(3) Peak table generation.

(4) Contour decomposition. Approximates the spectrum as a sum of Gauss or Lorentz contours.

NMR Spectrum Processing. (1) Calculation of transition frequencies and their intensities from chemical shifts and *JJ* coupling constants using Hamiltonian diagonalization. The simulated spectrum is presented using the spectrometer frequency and half-width of the Lorentz contours. The values of chemical shifts and *JJ* coupling constants may be assigned to atoms in a structure.

(2) Periodic table with relative nucleus frequencies, their spins, the natural abundance of isotopes, and relative sensitivities.

(3) Assignment checking. Compares values of the chemical shifts assigned to an atom(s) with that defined in the training set. The procedure was described in detail earlier.[14]

(4) Verification of the chemical shift axis direction. The procedure was described earlier.[14]

(5) Interconversion of chemical shifts using different NMR standards.

JCAMP IR and NMR Files. Reads the spectrum and auxiliary information (spectrometer frequency, sample preparation, etc.) from a JCAMP file. Any kind of spectral encoding—TABULAR DATA, PACKED, SQUEEZED, DIFFERENCE, and DIFDUP forms[9]—can be read. FID can also be read. Both fast Fourier transformation[15] and phase correction are available.

3.4. ChemSQL. ChemSQL is a service DLL. It enables the data exchange with the SQL server and allows one to browse/edit the data stored on the SQL server. Using SQL technology, a database can be edited from different locations at the same time. The Borland database engine (BDE) approach is used for developing this application. Ideally, using BDE, the same commands are used for different servers. In reality, slight modifications are required to create the client application for a selected SQL server. Three SQL servers—Oracle, MS SQL, and Interbase—can be used for data exchange with ChemSQL. All fields listed in Table 1, except for database link and WWW link, can be stored at SQL servers. Spectral information is also accepted. *Real*, *integer*, *string*, *text*, *Boolean*, *date*, and *time* fields, if they have neither *range* nor *atom assigned* nor *array* flags, are stored as native server variables. Other fields, inspite of having *range* or *assigned to atoms* flags, are stored as BLOB records. For application developers, ChemSQL contains ActiveX controls to display and edit BLOB fields. Also, an in-process OLE automation server is included to get BLOB data as WIN API HEnhancedMetafile and HBitmap in memory structures. The Bitmap or Metafile pictures can be inserted into custom reports.

Tables in a SQL database are created prior to execution of data import. ChemSQL contains a command for the creation of tables with the appropriate field types. In the next step it is necessary to match the fields in CheD and in the SQL databases. Generation of screens,[16] which are used for

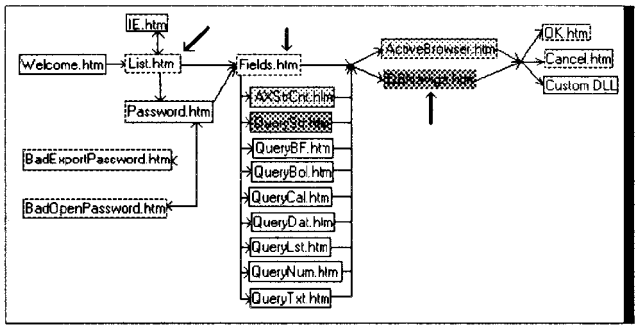

Figure 3. Web scenario of chemical data exposure.

substructure search acceleration, is made by the selection of a SDF file containing a list of screens.

3.5. ChemWSer. Chemistry applications of the Internet have grown significantly during recent years.[17,18] A number of chemical databases are available on the Web. Structure and substructure search engines are used to retrieve chemical data. All of the above-mentioned servers are unique: they are located at individual sites used as a part of a WWW data presentation by their owners. ChemWSer is packaged as a distributive application. It can be installed on Windows NT/95/98 computers within minutes.

ChemWSer consists of three applications: the setup program, the administrative tools, and the ISAPI DLL. In addition, a number of ActiveX controls are used for building the data presentation—query. ChemWSer Administrator is an application which allows one to expose or to hide databases on the Internet, to make links to HTML documents, and to define DLLs for session termination (see below). ISAPI DLL receives commands from a Microsoft or Netscape Web server and gives a response to it. Using ISAPI DLL instead of a CGI application has a number of advantages: (1) initialization is performed only once, during the first call to the DLL; (2) the current user session parameters can be stored in memory and be available while browsing different HTML pages. ISAPI DLL performs three tasks (Figure 2).

(1) It receives and sends data to the CheD application. CheD being a client, an internal (not HTML) format is used for the data exchange. A number of services can be executed: search in databases, sort, and selection of data by a list of values. The Search command is executed in the background: when the first record is found, it is immediately displayed to the client, and the search continues on the server. As new records are found, they also become accessible.

(2) It exposes the databases via the WWW. Two types of sessions, which are called Active and HTML, can be executed (Figure 3).

(3) ISAPI DLL is a JPEG image source for the stored data—chemical structures, spectral data, OLE objects, etc.

The Web scenario is shown in Figure 3. The thick arrows show possible entry points. A number of databases can be exposed, and are presented to the client as ≪List.htm≫. The client selects the database and the type of session—HTML or Active—from this page. If the selected database has passwords for opening or exporting data (see above), the client has to submit them. Confirmation of the password for data export is required only for an Active session, wherein the client can download a database and save its contents in a SDF or JCAMP file. The next step is to select the type of

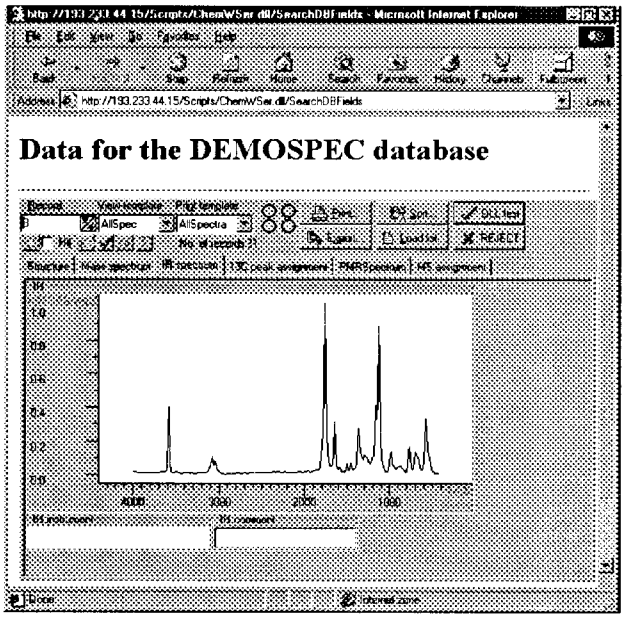

Figure 4. ActiveX session data presentation.

data retrieving—either a field to *Search* or the *Retrieve all* command with the Fields.htm page. One can insert a reference directly into this page as an HTML file; password confirmation, if required, will be executed. If a field for a search is selected, the query builder is presented. What type of query builder depends on the data type stored in the selected field. The query builder for a chemical structure search differs for HTML and ActiveX sessions. For an HTML session a Molfile with the query has to be created using any available structure editor and the content has to be inserted into the multiline text control. Otherwise an ActiveX control is loaded for query structure building.

The database content is shown on the ActiveBrowser.htm for an ActiveX session and the DBNaviga.htm for an HTML session. The ActiveBrowser.htm contains an ActiveX control (Figure 4), which receives binary data from ISAPI DLL. It should be noted that the data are not pictures; for example, one can zoom and scroll the spectrum, save it in JCAMP format, etc. Print and Display forms for data presentation can be selected. In an HTML session (Figure 5) data are shown as JPEG images. It is possible to make a direct jump to the query definition page or to browse the database content.

Besides the selection of Search for Data, the client may be used manually to select records by clicking the "Hit" checkbox both in HTML and in ActiveX sessions. The final selected records are used for the continuation of a session with a custom-written DLL (see below). All HTML pages are generated dynamically. Because ChemWSer is distributive, pages are generated from templates. Templates contain reserved keywords, which are substituted by ChemWSer. Termination of a session with ChemWSer can be customized also. Two ways to terminate are the following.

(1) Two (or one) HTML pages might be defined for a database. For example, they might be called "OKPage" and "CancelPage". References to these pages are defined during the administration of ChemWSer. It is possible to make reference to a single page or to not define a reference. A reference includes a page location and a button caption.

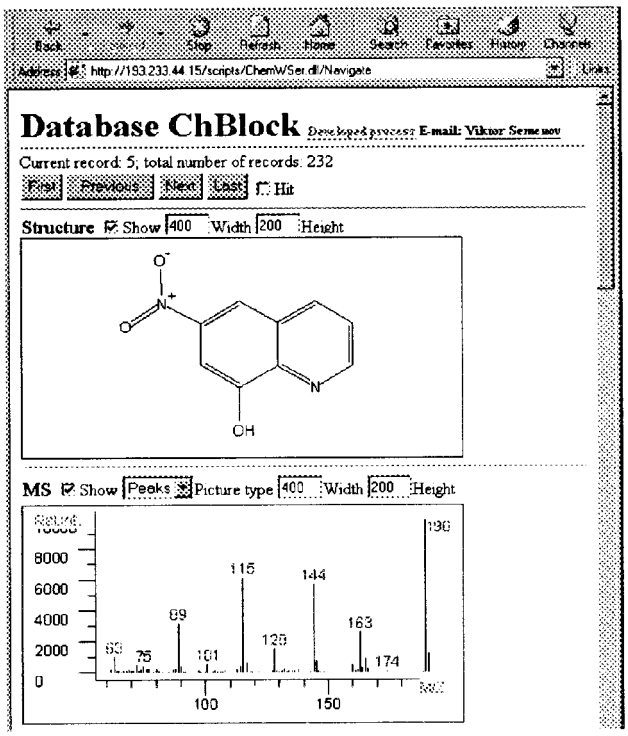

Figure 5. Data presentation in the HTML session.

When the client looks for data, these buttons are visible on the HTML page (or ActiveX control). Clicking on them will cause a jump to the referenced page.

(2) A custom-written DLL might be assigned to a database. The DLL should have a method which is exported by the name "OKResponce". The method has two parameters: a filename, where identifiers of records selected by the client are stored, and an output buffer. The buffer has to be filled with text in HTML format. The session customization by DLL gives administrators much more flexibility. For example, the DLL may call the SQL server and select the data with identifiers, stored in a file. Then, for example, the DLL can provide the client with a list of articles or calculated prices of selected compounds or make an invoice.

3.6. Implementation. Programs were tested on different computers with Windows95, Windows98, and Windows NT 4.0 operational systems. Tests of time of command execution, described here, were performed on a Pentium II processor, 450 MHz speed, 128 MB of RAM, Windows 98. The Delphi 3 Client-Server Suite (Borland Corp.) was used for the development of the programs. Personal Oracle 7 for Windows 95 (trial version), Oracle 7 Workgroup Server for NT (trial version), Interbase SQL server 4.2 for Windows 95/NT, and Microsoft SQL server version 6.50 were used for testing ChemSQL. Microsoft Personal Web server 4.0 (Windows 98) and Microsoft Internet Information Server 4.0 (Windows NT) were used for the ChemWSer testing.

The minimal computer configuration to run CheD is an Intel 486 100 MHz processor, 16 MB of RAM, Windows 95. Desirable configuration: Pentium processor, 32 MB of RAM. Additionally 24 bytes/ record in the database is desirable. For example, if one manipulates a database with 1 000 000 records, a computer with 256 MB of RAM should be used for the access.

Demo versions of the software described are available at http://ched.ipac.ac.ru. Examples of data display via the WWW can be found at http://www.chemical-block.com/read/welcome.htm.

ACKNOWLEDGMENT

We express our gratitude to Dr. Alexander Kutin, Dr. Andrew Podgursky, and Dr. Bogdan Ugrak (Zelinski Institute of Organic Chemistry, Moscow) for discussions and program testings and Dr. Vladislav Kuusk (University California, San Francisco) and Dr. Paul Stott (Uniroyal Chemical) for manuscript revision.

REFERENCES AND NOTES

(1) Henderson, K. *Client/Server developer's guide with DELPHI 3*; SAMS Publishing: Indianapolis, IN, 1996; p 730.
(2) Warr, W. A. Communication and Communities of Chemists. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 966−975.
(3) Product information may be obtained from Molecular Design Ltd., San Leandro, CA 94577 (http://www.mdli.com).
(4) Product information may be obtained from Cambridge Soft Corp., 100 Cambridge Park Dr., Cambridge, MA 02140 (http://www.chemoffice.com).
(5) Brecher, J. S. The ChemFinder WebServer: Indexing chemical data on the Internet. *Chimia* **1998**, *52*, 658−663.
(6) Williams, A.; Mityushev, D.; Shilay, V.; Kvasha, M. *An Integrated software system for Spectral Management for NMR, MS, IR, and UV-Vis and Chemical Structures*; Advanced Chemistry Publications: Toronto, Canada, www.acdlabs.com/publish.
(7) Hamilton, W. C. *Acta Crystallogr*. **1965**, *18*, 502−509.
(8) Dalby, A.; Hourse, J. G.; Hounshell, W. D.; Gurchurst, A. K. I.; Grier, D. L.; Leland, B. A.; Laufer, J. Description of several chemical structure file formats used by computer programs developed at Molecular Design Limited. *J. Chem. Inf. Comput. Sci*. **1992**, *32*, 244−255.
(9) McDonald, R. S.; Wilks, P. A., Jr. JCAMP-DX: a standard form for exchange of Infrared spectra in computer readable form. *Appl. Spectrosc.* **1988**, *42*, 151−162.
(10) Gasteiger, J.; Hendriks, B. M. P.; Hoever, P.; Jochum, C.; Somberg, H. JCAMP-CS: A standard exchange format for chemical structure information in computer-readable form. *Appl. Spectrosc.* **1991**, *45*, 4−11.
(11) Bebak, H.; Buse, C.; Donner, W. T.; Hoever, P.; Jacob, H.; Claus, H.; Pesch, J.; Roemelt, J.; Schilling, P.; Woost, B.; Zirz, C. The standard molecular data format (SMD) as an integration tool in computer chemistry. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 1−5.
(12) Berger, D. J. ChemWindow DB (4.0). *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 409−409.
(13) Holliday, J. D.; Ranade, S. S.; Willett, P. A fast algorithm for selecting sets of dissimilar molecules from large chemical databases. *Quant. Struct.-Act. Relat*. **1995**, *14*,501−506.
(14) Trepalin, S. V.; Yarkov, A. V.; Dolmatova, L. M.; Zefirov, N. S.; Finch, S. A. E. WinDat: An NMR Database Compilation Tool, User Interface, and Spectrum Libraries for Personal Computers. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 405−411.
(15) Cooley, J. W.; Tukey, J. W. An algorithm for machine calculation of complex Fourier series. *Math. Comput.* **1965**, *19*, 297−301.
(16) Heller, S. R. Chemistry on the Internet-the road to everywhere and nowhere. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 205−213.
(17) Wiggins. G. Chemistry on the Internet: the library on your computer. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 956−965.
(18) Murray-Rust, P.; Rzhepa H. S.; Whittaker B. J. The Word Wide Web as a chemical information tool. *Chem. Soc. Rev.* **1997**, *27*, 1−10.

CI000039N