Geological Society of America
Special Paper 397
2006

Building the American Mineralogist Crystal Structure Database: A recipe for construction of a small Internet database

Hareesh Rajan
Department of Computer Science, University of Arizona, Tucson, Arizona 85721-0077, USA

Hinako Uchida
Deborah L. Bryan
Department of Geosciences, University of Arizona, Tucson, Arizona 85721-0077, USA

Ranjini Swaminathan
Department of Computer Science, University of Arizona, Tucson, Arizona 85721-0077, USA

Robert T. Downs
Michelle Hall-Wallace
Department of Geosciences, University of Arizona, Tucson, Arizona 85721-0077, USA

ABSTRACT

Crystal structure data represent one of the most important resources for developing scientific knowledge and should be archived in ways that make them easy to access and preserve. The American Mineralogist Crystal Structure Database currently contains every crystal structure published in *American Mineralogist, The Canadian Mineralogist, European Journal of Mineralogy,* and *Physics and Chemistry of Minerals*. It is maintained by the American and Canadian mineralogical societies and is freely accessible through the Internet. The database consists of the data, server-side search and retrieval software and user-side analysis software. It is managed through a partnership of PHP and MySQL programming that provide dynamic construction of Web pages and search procedures. The purpose of this paper is to describe the database and its implementation and to illustrate how to construct similar small, interactive Internet databases.

Keywords: crystal structure, database, visualization.

INTRODUCTION

The literature in many disciplines is full of small data sets created through individual studies, which over decades and even centuries have defined the core of understanding for the discipline. These data sets are fully recoverable by an experienced individual researcher, but the recovery often requires significant amounts of time and effort. The move to digital scientific journals is one step toward making current data sets more readily available, as is the creation of many new digital libraries. We have created a comprehensive database of all crystal structure data published in the journals of the Mineralogical Society of America, the Mineralogical Association of Canada, and the European societies through a time-consuming yet richly rewarding process. The database not only provides specific information about any one of 8000 different crystal structures in one location, but it also

Rajan, H., Uchida, H., Bryan, D.L., Swaminathan, R., Downs, R.T., and Hall-Wallace, M., 2006, Building the American Mineralogist Crystal Structure Database: A recipe for construction of a small Internet database, *in* Sinha, A.K., ed., Geoinformatics: Data to Knowledge: Geological Society of America Special Paper 397, p. 73–80, doi: 10.1130/2006.2397(06).

provides freely available analysis software that is fully integrated with the database's data retrieval system.

The American Mineralogist Crystal Structure Database consists of (1) a collection of crystal structure data, (2) an Internet site that allows search and retrieval of data sets using a browser, (3) server-based PHP: Hypertext Preprocessor (or PHP) and FTP (file transfer protocol) software to provide direct access to the data using PHP queries, and (4) user-based software that provides for interactive analysis. The database can be found at http://www.geo.arizona.edu/AMS/amcsd.php or from the Web sites of the mineralogical societies. In this paper, we describe the unique characteristics of the data and their storage, development of the database, and the tools for integrated data retrieval and analysis. Our goal is to describe the functionality of the American Mineralogist Crystal Structure Database and to provide information on how to construct a similar database and Web interface. Downs and Hall-Wallace (2002) provide a detailed description of the purpose of the database and its user-based analysis software.

The structure of a crystal represents a minimum energy configuration adopted by a collection of atoms at some set of conditions. In principle, all the physical and chemical properties of any crystalline substance can be computed from its crystal structure. The determination of crystal structures and the deduction and understanding of these computational algorithms constitutes a major part of scientific research in physics, chemistry, biology, medicine, mineralogy, geology, and material sciences. As such, crystal structure data represent one of the most important resources for developing our scientific knowledge and thus should be archived in ways that make them easy to access and preserve. However, these data are often cumbersome to retrieve and verify from the literature, and even more difficult to analyze without a strong background in crystallography. Thus, the National Science Foundation, in cooperation with the Mineralogical Society of America, the Mineralogical Association of Canada, and the University of Arizona, has established a collection of mineralogically important crystallographic data sets that are freely accessible through the Internet. This database is an important resource to both of these mineralogical societies, and they are committed to maintain it as part of their outreach programs.

GATHERING AND PREPARING THE DATA SETS

The database contains every experimentally determined crystal structure reported in *American Mineralogist*, *The Canadian Mineralogist*, *European Journal of Mineralogy*, and *Physics and Chemistry of Minerals*. At present there are about 8000 individual data sets representing about 7000 unique mineral and chemical phases. Data from *Acta Crystallographica* are now being added. The data represent structures determined at ambient conditions as well as at temperature or pressure. Assembling the database is a multistep process that includes:

- examining each volume of the journals to identify all papers that report crystal structures;
- manually entering the reported data in the database;
- computing various crystal chemical parameters and verifying the consistency of these parameters with those that are reported;
- contacting the authors about irresolvable inconsistencies between reported and computed parameters; and
- incorporating comments when changes are made to the originally published data.

Each record in the database consists of a bibliographic reference, cell parameters, symmetry, and atomic positions, as well as displacement parameters and site occupancies when applicable. An example of retrieved data sets is provided in Figure 1.

The first part of each data set contains identifying information, bibliography, and notes, while the second part of a data set contains the crystallographic parameters. The first line of a data file contains an identifier, such as the name of the mineral or formula of the chemical species. The next lines contain the names of the authors, followed by the journal reference, title of the paper, and any additional notes. The crystallographic data begin with a listing of the cell parameters and space group. The rest of the data set is a listing of the atoms, their positional and displacement parameters, and occupancies. A header is provided that defines right-justified columns. The name of each atom identifies the occupying elements, with additional identifiers added when appropriate. For instance, "Oco" identifies a particular oxygen atom in the albite structure. Each data set is stored in a separate and uniquely named ASCII file with the extension .amc and is put in a read-only directory of an FTP server.

Quality control is a major concern, and considerable effort is expended to ensure that the data are correct. Each data set is verified for accuracy and omissions before being added to the database. We use METRIC, software developed by Bartelmehs et al. (1993) at Virginia Tech, to compute the crystal chemical parameters that are typically found in many crystal structure publications. The crystal chemical parameters include bond lengths, angles, polyhedral volumes, distortion indices, and such, as recently reviewed in Hazen et al. (2000). The crystal chemical parameters calculated by METRIC are compared with those included in the original publication to provide a check of the internal consistency of the data. If a discrepancy is found, then we try to determine its source and correct it. Many errors are typographical, while others are more complex. The authors are contacted to correct an error, if possible. The database contains only the corrected data set.

Developing a fast and efficient search engine requires the data to be well organized. Storing all the data sets within a large single file was initially considered but rejected in order to simplify the transfer of individual data sets between our server and a user's computer. Hence, the data sets are stored and maintained in individual files, with each file containing the data for a single crystal structure. Individual data files are simple to manage and edit. More importantly, in the event of unforeseen needs, the format of a single data file can be easily modified. The data sets are stored as individual ASCII files, averaging ~20 lines each. It is anticipated that the final set of data will consist of ~50,000 files.

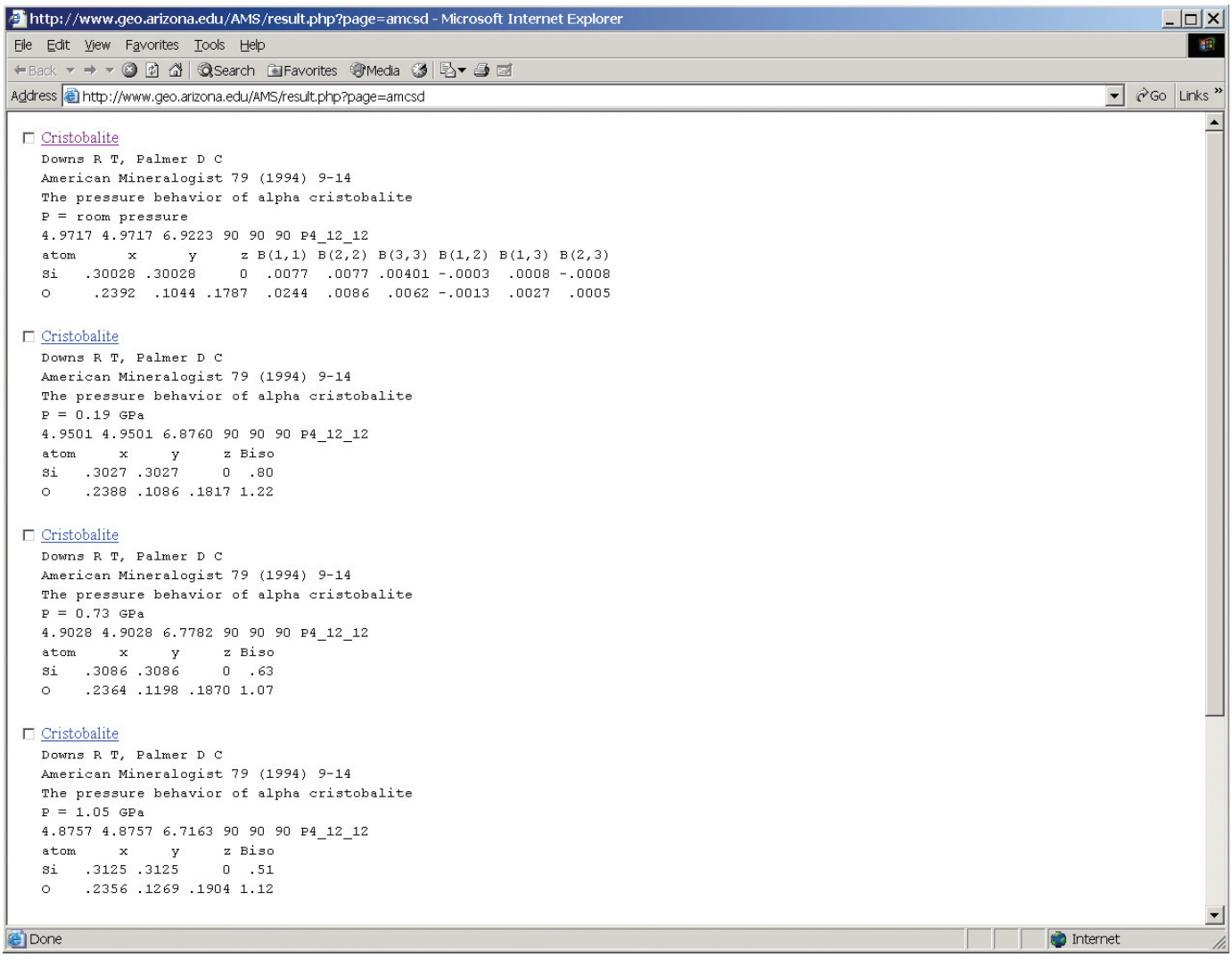

Figure 1. This example of a retrieved set of data for cristobalite illustrates the format of the data. Four data sets are displayed in the figure. The underlined mineral names indicate FTP hyperlinks that provide automatic launching of user-defined software. Check boxes are provided for each data set to permit the download of a single file that contains several selected data sets.

THE VISUALIZATION AND ANALYSIS SOFTWARE

Crystallographic parameters are the foundation of research in material properties, and being able to visualize and manipulate crystal structures is fundamental to understanding them. A suite of Windows®-based visualization and analysis tools, fully integrated with the database, has been developed to make the database more useful as a scientific and educational tool. Three key applications are (1) XtalDraw, an application that creates graphical representations of a crystal structure, (2) METRIC, described above, and (3) XPOW, an application that computes and displays a theoretical powder diffraction pattern of the crystal. The software currently only operates in Windows®.

When a crystal structure in the database is accessed, its image can automatically be displayed with the XtalDraw application (Bartelmehs et al., 1993; Downs and Bartelmehs, 1996; Hazen and Downs, 1996; Downs, 1998; Hazen et al., 2000). Currently, this stand-alone software contains options to draw the crystal structure with ball-and-stick, polyhedral, and thermal ellipsoidal renderings (Fig. 2A). The user can rotate the image with the arrow keys, or manually enter directions to view in direct or reciprocal space coordinates. Users can expand or shrink the image, add or delete atoms from the field of view, change the colors of atoms, their sizes, the bonds, and so forth. The software produces publication-quality bitmaps. Carefully selected default viewing parameters are one very powerful feature of XtalDraw that make it easy to use. For instance, after a data file is read, the structure is initially displayed in a default orientation established by the International Tables for Crystallography. The atoms are drawn in a set of colors suggested by Lipson and Cochran (1957)

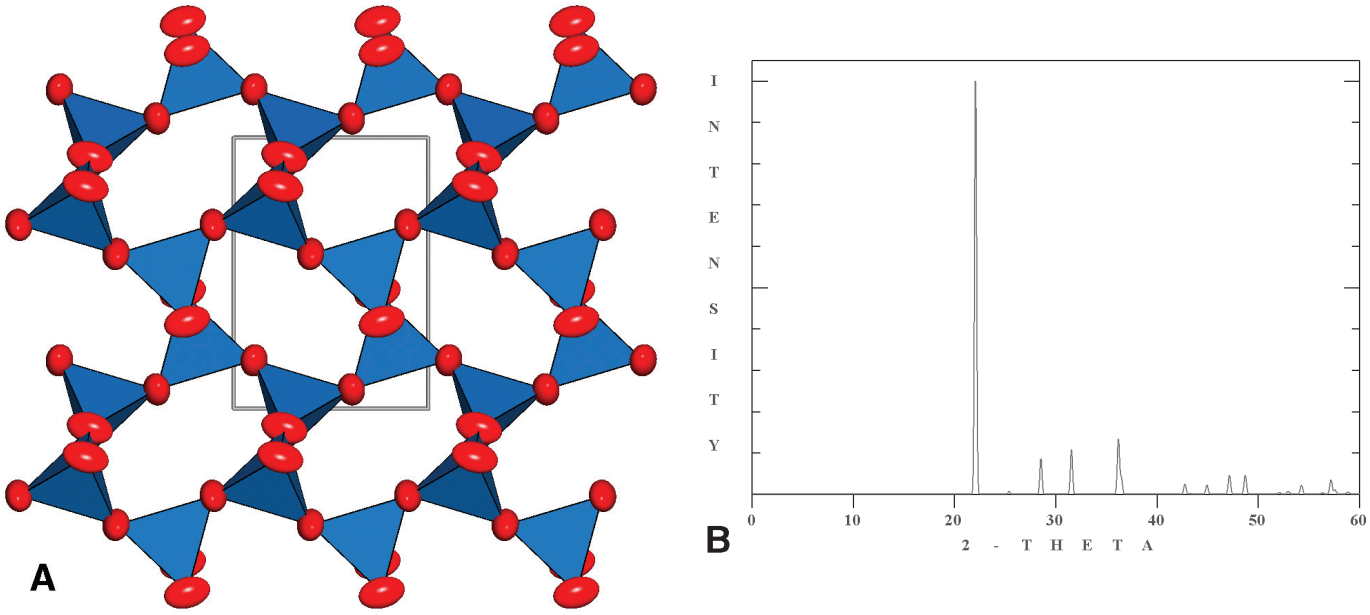

Figure 2. Two images that illustrate the output produced by the accessory software. A: A mixed polyhedral and thermal ellipsoid rendering of the crystal structure of cristobalite produced by XtalDraw. B: A simulation of an X-ray powder diffraction pattern computed for the structure of cristobalite by XPOW.

to represent elemental species, and in sizes that scale to the Shannon and Prewitt (1969) radii. Atoms can be classified as cations or anions, and the coordination sphere is automatically estimated for polyhedral representations. Users have the option to change these settings at any time and to establish their own default settings. As a feature of XtalDraw, the METRIC software can be launched to calculate many of the important crystal chemical parameters that are used by today's researchers, such as bond lengths and angles, polyhedral volumes and distortion parameters, and vibrational amplitudes and orientations (Hazen et al., 2000).

Another important property of a crystal that can be computed from its structural data is its diffraction pattern. Neutron and X-ray powder diffraction patterns, generated from a fixed source or from an energy-dispersive source, such as at a synchrotron, can be computed with the software XPOW (Downs et al., 1993). XPOW can calculate the diffraction pattern for any crystalline substance and provide an interactive display of the pattern, as it would be generated by a conventional powder diffractometer (Fig. 2B). The user can alter the radiation source wavelength, the peak widths, limits, and orientation of the display. It is also of great use to identify unknown materials by providing a complete diffraction pattern, and not just the few d-spacings listed in a search-match table.

By using the FTP hyperlink, discussed in greater detail below, the software can be automatically launched by a visit to the site and selection of a data set. The browser or operating system must be configured to automatically associate .amc files with the application. The simple structure of the data sets and database also allows any other software application capable of receiving instructions through "command line arguments" (as is typical in drag-and-drop applications) to open and manipulate the data.

DATA RETRIEVAL METHODS

A good database should be comprehensive, accurate, and easily analyzed; it must also be easily accessed, and data must be easily retrieved. Currently, access is provided through three methods: (1) a browser interface with an interactive search procedure, (2) retrieval by sending a PHP query directly to the database without use of the Web search procedure, and (3) direct FTP from an anonymous FTP site.

The interactive search procedure consists of an HTML page (Fig. 3) that allows the user to specify several criteria, including the names of the minerals or chemical species, the authors and titles of the relevant journal articles, mineral chemistry, unit cell dimensions, and space group symmetry. There is also a tool to search for arbitrary words or phrases. Furthermore, search criteria can be combined with the logical conditions *or* and *and*. The interface is constructed dynamically using PHP programming, and the search is conducted with the database management software MySQL. We use an Apache Web server running on a Linux platform, which is well suited for PHP and MySQL.

After the search criteria are defined and the search is initialized, the PHP program calls MySQL, which returns the names of appropriate .amc data files. A search may result in a single data set or multiple data sets. The PHP program collates the contents of the matching data files and displays them in a new browser window. Each of the data sets has an FTP hyperlink

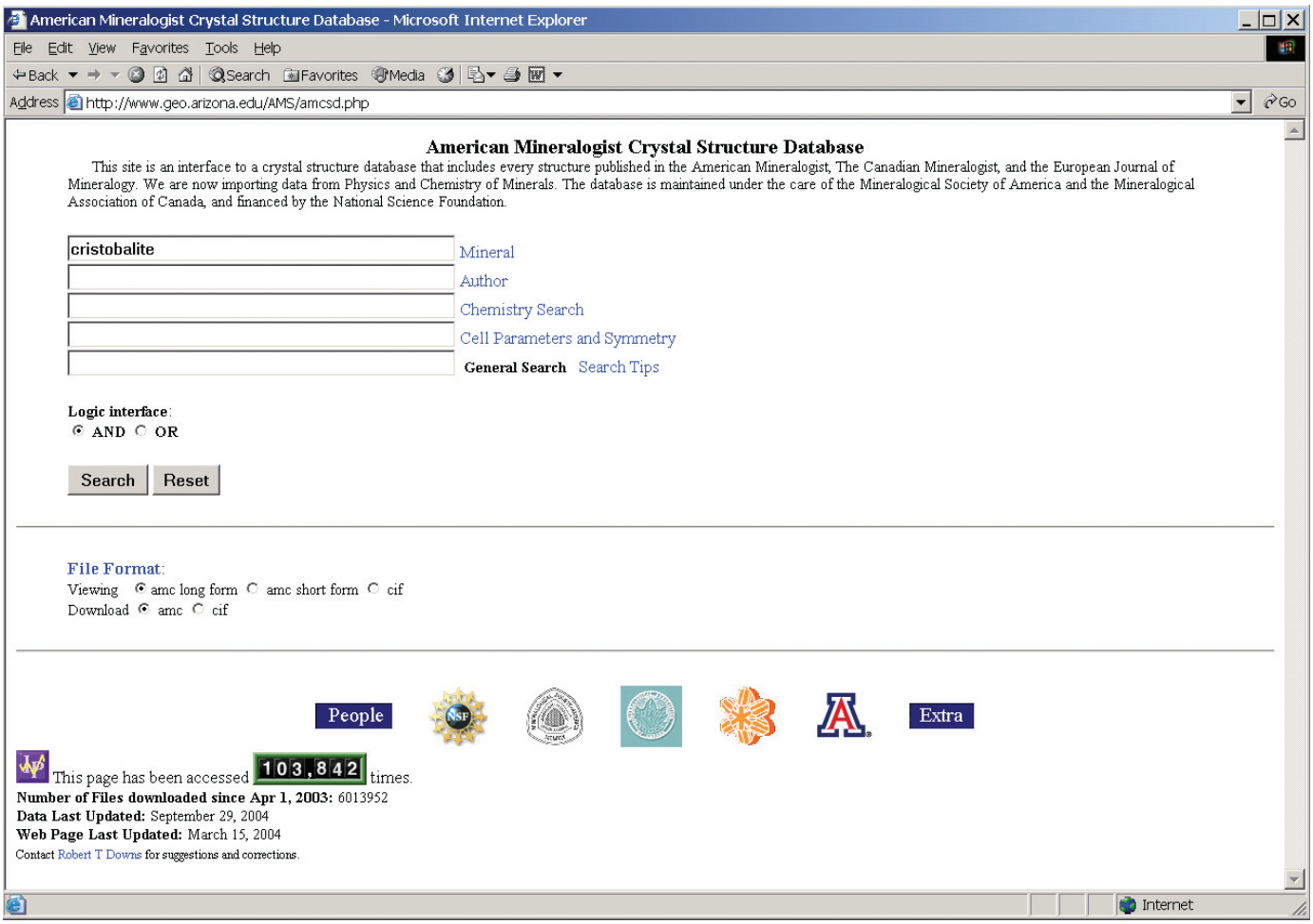

Figure 3. The American Mineralogist Crystal Structure Database Web page. This is the interface that can be used to search for crystal structure data files. The HTML code for this page is created dynamically with PHP programming. In this image, the mineral name "cristobalite" was been typed in the edit box. Clicking the *search* button will retrieve the data sets for cristobalite and display the data as shown in Figure 1.

associated with it. To retrieve the data, the user can (1) cut and paste the relevant text from the browser into another application, or (2) select the FTP hyperlink(s) and download the entire set of matching data or any subset of it. The FTP hyperlink has the form, for example,

Cristobalite.

When activated, this hyperlink automatically transfers the data file from our server to the user's computer. The user is relieved from the trouble of directly dealing with the transfer. Furthermore, the FTP link provides the ability to launch an application directly on a user's computer simply by clicking the link. If the file name extension of the data, .amc, has been associated with an application on the user's local computer, then the application will be automatically launched and will open the file. For instance, XtalDraw and XPOW can be launched in this manner. This feature is modeled after the .pdf interface with Adobe Acrobat Reader®.

In addition to accessing data through the interactive search procedure, sending a PHP query directly to the Web server can initialize a search. A successful search brings up the data, as in Figure 1, without going through the search interface Web site. This method for accessing the data is currently being employed by databases such as that found at http://www.webmineral.com and is also a good way to provide access to the data through educational Web pages that need to retrieve a specific file. Access is initialized through a hyperlink that launches a PHP query, for example,

http://www.geo.arizona.edu/AMS/result.php?mineral=cristobalite.

This command produces a search for cristobalite using the *mineral* table in the database. The details for constructing searches of this type are given at the Web site.

Users can also gain direct access to the entire database, through an FTP connection to our anonymous FTP site. This enables the secure transfer of files between two computers and is

often more efficient when large numbers of files are being transferred. In general, FTP software is more cumbersome to use than a browser, thus we also incorporated the FTP hyperlinks, discussed earlier, into our interactive search procedure.

Appropriate security measures are taken to prevent malicious users from damaging the data files. These include constraining the storage directory as the only directory on the server to have public access, and defining this directory and the file attributes as "read only." Users are not permitted to upload their files into our storage directory. These security issues are operating system dependent and must be addressed on an individual basis.

DESIGNING AND MANAGING THE DATABASE

We use MySQL, fast and reliable database management software, combined with the PHP scripting language to create our database structure, to populate and maintain the database, and to support data retrieval. These are two of the most widely accepted tools for Internet database applications. MySQL runs on all platforms and performs exceptionally well for small databases such as ours. Furthermore, it is open-source and hence cost-effective.

The type of data we accommodate and the requirements of the interactive search procedure guide the database design and management. The database must associate data set file names with the search criteria, which include *minerals*, *authors*, *titles*, *chemical elements*, *cell parameters*, *space group symmetry*, and a general *keyword* search. For each searchable criterion, we have created a table using the MySQL *Create* command. Each table is mainly composed of individual records containing the entity identifier in one field and a corresponding data set file name in another. For example the *minerals* criteria table contains records with a mineral name (e.g., cristobalite) in one field, and a file name containing the crystal structure data set for that mineral (e.g., 01553.amc) in another field. The *authors* criteria have a table with records that contain a field with an author's name and another field with a file name containing the associated data set. The database may contain several different data sets for each mineral or for each author, and each one of these data sets constitutes a separate record in the table.

The MySQL *Create* command constructs the table but does not put any data into it. PHP scripts read the individual data files, extract the corresponding entities, and insert them into the corresponding criteria table. The *minerals* criteria table thus contains all the minerals present in the data files along with the names of their associated files. The *chemical elements* table contains all the atoms present in each of the crystal structures along with their associated file names. Similarly, the table for *cell parameters* and *space group symmetry* contains the values for a, b, c, α, β, γ, and *space group*, along with their corresponding file names. Our database undergoes regular updating because new data sets are continually being added. The PHP scripts read all the data files and rebuild the entire MySQL database during each update.

Once the tables are created and the appropriate data inserted into them, the database is complete and ready for query purposes. Data are retrieved and viewed in a variety of ways, using the MySQL *Select* commands. For instance, a simple query

> **Select** filename **from** minerals **where** mineralname = "Cristobalite"

will retrieve all the file names associated with the mineral "cristobalite" from the *minerals* table. Other queries are performed using the refinements of the *Select* commands, including *Order By* (sorting routines) and *Distinct* (removes duplicate entities) clauses.

The method of searching for cell parameters is a little different than for the other fields and requires different handling in the database design. The cell parameters are stored as real numbers rather than as text strings. The table for the cell parameters has seven different fields: one for file name and six others for the six different cell parameters. A user searching for cell parameters usually would not know their exact values, and instead could only provide an estimate. The search criteria are established through a Web page as shown in Figure 4. MySQL supports searches for real numbers that lie within a given range or tolerance. This option is used for searching for cell parameters and is invoked with a command such as

> **Select** filename **from** cell_parameters **where** $a >= 4.97$ **and** $a <= 4.99$.

MANAGING THE WEB INTERFACE THROUGH PHP

The user interface to the searchable fields includes interactive HTML elements such as text boxes, drop-down list boxes, radio buttons, push buttons, etc., created with the server-side scripting language PHP. More importantly, PHP securely supports the MySQL database. It is the link between the database and the Web page.

The Web search interface is dynamically created with PHP using the *Select* commands to access the criterion tables in the MySQL database. The results of the queries populate the search criteria fields in the HTML form. The user is presented a detailed listing in a drop-down list box of all the *minerals*, *authors*, *titles*, and *space groups* available in the database. The *chemical elements* search provides a periodic table from which the user can select elements to constrain a search (Fig. 5) and is described in detail below. The *cell parameter* and *space group* search is done through another Web page illustrated in Figure 4 and described above.

Each of the different search tools is used to create search requests that are processed using scripts written in PHP, which provides numerous built-in functions to access the MySQL database. Any interaction with the database requires PHP to establish a connection to the MySQL server using the *mysql_connect ()* function. Once connected, our database is selected using the *mysql_select_db ()* function. All the queries in the script are subsequently directed to the database. Database queries are made using the *mysql_query ()* function, which creates SQL statements to be executed by the MySQL server. The results of the query are retrieved using the *mysql_fetch_row ()* function.

The Web server, on receiving a search request from the user, processes the PHP scripts to connect, select, query, and fetch the

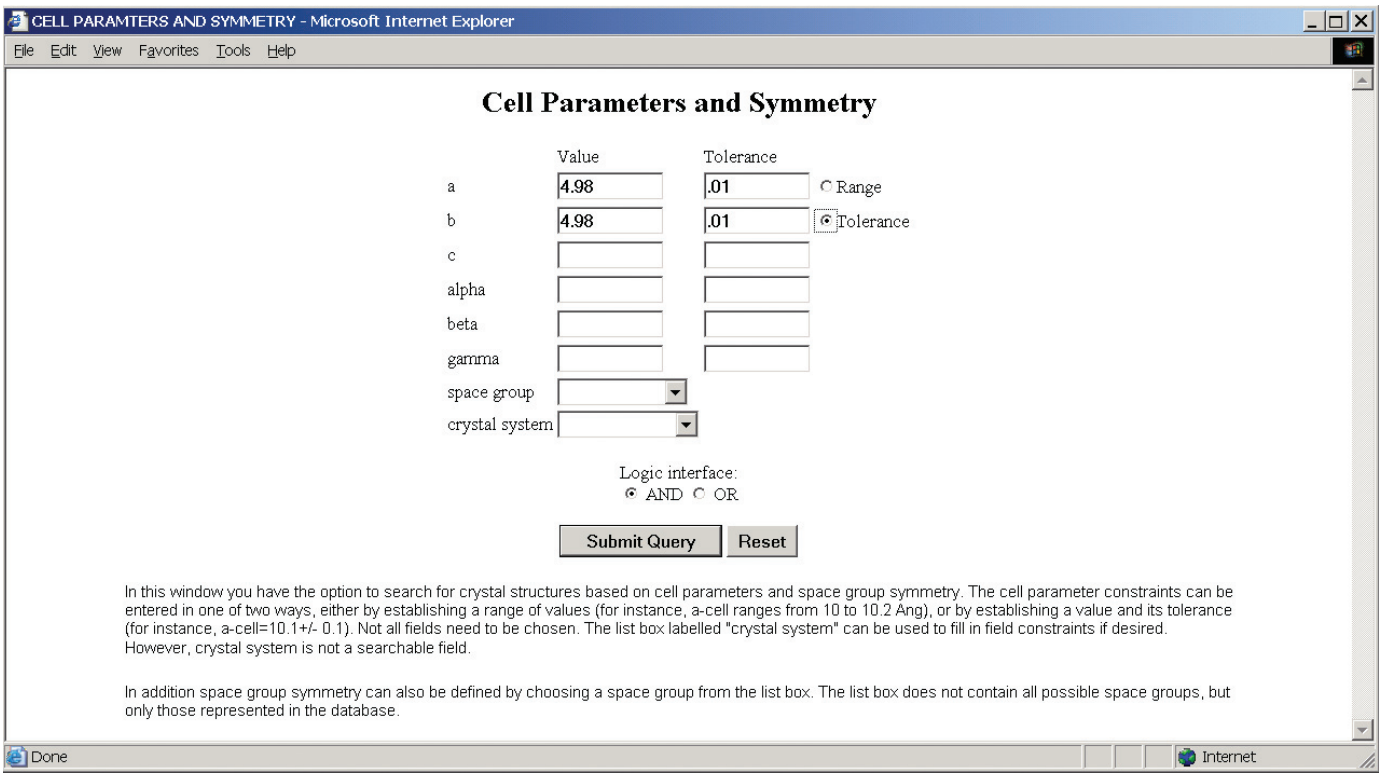

Figure 4. The Web page that is used to define cell parameters and space group symmetry. A search for cell parameters is handled in a different way than a search for the other fields because cell parameters are real numbers that normally can only be estimated by a user. Thus, the user defines a search for cell parameters by providing a search interval, or by providing a median value with a tolerance.

data. The MySQL database responds by executing the queries and sending the names of the files containing the selected items in a PHP script. PHP stores the results in one or more variables and uses the *print()* function to display the data files in HTML format, as output to the user. The *Limit* clause of the MySQL *Select* command is used to define the number of data files displayed on a single page. We designed a single page to display up to 100 data sets at a time. Individual check boxes associated with the output of each data set allow the user to choose a subset of the retrieved data. The user can select all or a subset of the data by clicking the check boxes. Then, clicking the *process the data* button launches the FTP command to transfer the selected data sets (Fig. 1).

Even though handling the search for chemistry can be done in the same way as a search for *mineral name*, *author*, or *title*, the complexity of a chemical search requires a different interface. This is because a good chemical search routine must permit the selection of many different elements at a time, as well as the exclusion of many elements. Therefore, we constructed an interactive and user-friendly periodic table interface (Fig. 5) for the *chemical elements* search, again managed by PHP. Users make their selection of elements directly from the periodic table by clicking the elements. This is mainly designed with the help of CSS (cascading style sheets), which offer great flexibility in controlling the overall presentation of the HTML document and also significantly reduce the amount of coding required for formatting. Each chemical element has an associated button that records one of three conditions for the element: (1) Find data that include this element, (2) find data that do not include this element, and (3) ignore the presence of this element in the data set. For example, all silicate minerals can be obtained by clicking only Si and O. All other elements are, by default, in state (3) and thus can be present in a data set. To obtain only the SiO_2 silica polymorphs, all chemical elements other than Si and O must be put into the exclude state (2). Instead of clicking each and every element to put them into this state, an *exclude others* button is provided that defines all other elements to be in the exclude state.

Lastly, a general *keyword* search is conducted using the individual data set files rather than the database tables. The general *keyword* search is done using the familiar *grep* command of Unix, which searches through the directory of individual data files for the desired words or phrases and returns the name of the files containing it. The *grep* command can be invoked from PHP using suitable system functions. The general search requires the user to input either words or phrases in the text box, with the words being case-insensitive. Since the search request for the *grep* command comes from the user input, care is taken to prevent the user from tricking the system

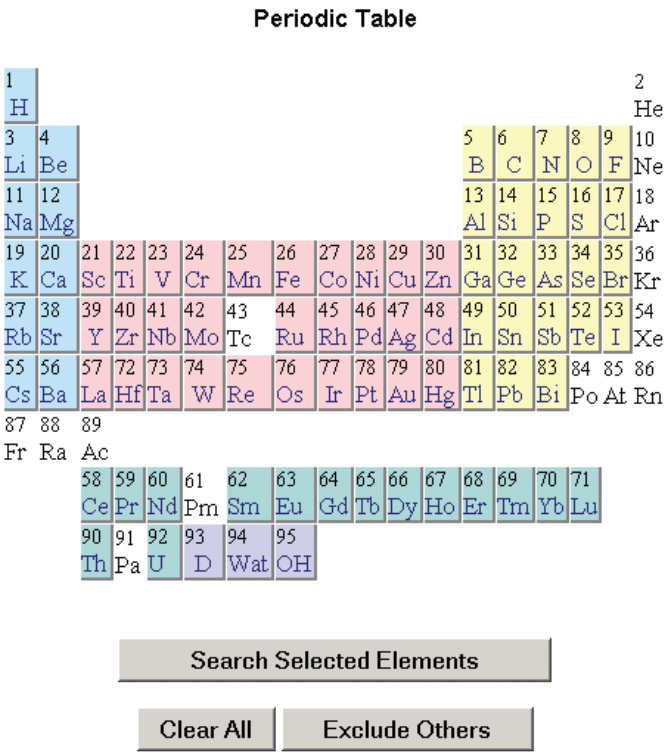

Figure 5. The Web page that is used to select chemical elements for the database search. It is created dynamically with PHP programming and cascading style sheets. Each element is in a table cell that behaves like a button and can be clicked into one of three different selection states.

into executing arbitrary commands. The *grep* utility is fast for searching, easy to implement, and fits our application's requirements. Other means for searching through a collection of files include search engines such as Web Glimpse or Google.

REFERENCES CITED

Bartelmehs, K.L., Gibbs, G.V., Boisen, M.B., Jr., and Downs, R.T., 1993, Interactive computer software used in teaching and research in mineralogy at Virginia Tech: Geological Society of America Abstracts with Programs, v. 25, no. 6, p. A-347.

Downs, R.T., 1998, Computer graphics simulation of compression mechanisms in crystals: IUCR-HP98, Argonne, Illinois, Abstracts, v. 21.

Downs, R.T., and Bartelmehs, K.L., 1996, Computer visualization of temperature and pressure effects on crystal structures: Eos (Transactions, American Geophysical Union), Spring Meeting Supplement, v. 77, no. 17, p. S261.

Downs, R.T., and Hall-Wallace, H., 2002, The American Mineralogist Crystal Structure Database: American Mineralogist, v. 87, p. 247–250.

Downs, R.T., Bartelmehs, K.L., Gibbs, G.V., and Boisen, M.B., Jr., 1993, Interactive software for calculating and displaying X-ray or neutron powder diffractometer patterns of crystalline materials: American Mineralogist, v. 78, p. 1104–1107.

Hazen, R.M., and Downs, R.T., 1996, Systematic crystal chemistry of high-pressure silicates: An interactive graphics demonstration: International Union of Crystallography XVII Congress and General Assembly, Seattle Washington, C-543.

Hazen, R.M., Downs, R.T., and Prewitt, C.T., 2000, Principles of comparative crystal chemistry, *in* Hazen, R.M., and Downs, R.T., eds., Comparative crystal chemistry: Washington, D.C., Mineralogical Society of America, Reviews in Mineralogy and Geochemistry, v. 41, p. 1–33.

Lipson, H., and Cochran, W., 1957, The determination of crystal structures: London, G. Bell and Sons, 345 p.

Shannon, R.D., and Prewitt, C.T., 1969, Effective ionic radii in oxides and fluorides: Acta Crystallographica, v. B25, p. 925–945.

Manuscript Accepted by the Society 18 June 2005

Geological Society of America
3300 Penrose Place
P.O. Box 9140
Boulder, CO 80301
(303) 447-2020 • fax 303-357-1073
www.geosociety.org