Automatic Spectral Database and Archive System for Optical Spectroscopy

D. J. KRAUS and R. H. FRENCH*

Central Research & Development Department, E. I. DuPont DeNemours & Co., Wilmington, Delaware 19880

With the advent of computerized data acquisition in spectroscopy, it has become possible to acquire large quantities of data with relative ease. Without a convenient method for archiving and accessing the data at a later date, a number of the capabilities inherent in these systems are unrealized. We have developed a generalized archival system which automatically archives all raw data and generates a rapidly searchable database of text information and spectral data for all spectroscopic results covering the spectral regions from 30 nm to 1000 μm (3.3×10^5 to 10 cm^{-1}). The system executes on a local area network (LAN) of personal computers and consists of a customized file-naming and comment convention that organizes data by research notebooks, an integrated data backup routine, and a spectral database system. The data backup and spectral database consist of two programs: (1) a DOS level copy routine to back up the system and data to an optical write once/read many (WORM) drive, and (2) a spectral database program written in the Lab Calc spectroscopy environment. Stored with each spectral library entry are searchable text fields of character data parsed from a standardized comment line which is entered prior to data acquisition. This system is used consistently with three different spectrophotometers, allowing all results to be treated equivalently for further numerical analysis.

Index Headings: **Computer applications; Infrared; Instrumentation, database; Spectroscopic techniques; UV-visible spectroscopy.**

INTRODUCTION

Archiving data on computers without a convenient, easy-to-use method of retrieval is hardly worth the effort. It may be easier and/or faster to actually recollect or reanalyze data than to sort through possibly hundreds or even thousands of files which have little more than a file name to access the data. Lab Calc† has provided the tools (the Search module) to develop an automatic archival method which is being used in our laboratory for optical spectroscopy of ceramics and optical materials from the vacuum ultraviolet (VUV) to the far-infrared. The end product of this procedure is an elegant archive library with high-speed searching ability for enormous amounts of spectral data. The system automatically archives all raw data to optical media, while maintaining a spectral database which can be searched for a text string imbedded in a database record amidst more than ten library files in under ten seconds, and emulate animated video viewing of spectra with over 400 files per minute being streamed by. The system has proven to us to be vital to the organization of data and extremely helpful in locating spectra.

DISCUSSION

The development of this technology was spurred by our ability to acquire high-resolution spectra over an exceedingly wide energy range, thereby producing large numbers of relatively large data files. Spectral data sets are acquired on three different instruments: a custom, laser-plasma-sourced, dispersive VUV/UV spectrophotometer,[1,2] with Vis/NIR spectra acquired with a λ-9 spectrophotometer,[3] and an FT-IR spectrophotometer.[4] The huge volume of data arises because the VUV spectra, for example, are being collected over the range of 700 to 28 nm with a 0.1-nm or 0.3-nm resolution. Raw data files are acquired in the various software packages particular to the spectrophotometer and then imported into the standard Lab Calc format. In the VUV/UV, the original raw data files are ~6 KB in size. For each spectral region, three imported raw data files (~4 KB ea.) representing dark current, lamp intensity, and sample signal are used to calculate the absolute reflectance or transmission of a sample. For production of the initial resultant file (~65 KB), 16 overlapping spectral regions (~4 KB ea.) are concatenated. The optical properties and dielectric constant are typically calculated from the reflectance file with the use of an FFT-based Kramers-Kronig analysis performed in Lab Calc,[5] producing more results (~65 KB ea.) en route to further analyses. In a single morning's work more than 1.1 MB of data are generated in over 140 files. In the past fifteen months over 1.5 gigabytes of data have been processed in this lab with more than 3000 calculated files entered in various spectral database libraries. In Fig. 1, a typical spectral database entry, consisting of the full resolution spectrum and an associated database text page, is shown for file K0509M, the room-temperature optical conductivity of $MgAl_2O_4$ from 5 to 40 eV.

As shown in Fig. 2, the hardware system layout, a local area network (LAN) of high-performance microcomputers is used for this spectroscopy, with three computers devoted primarily to data acquisition, after which the data are transferred via the network to four additional computers for further analysis. These computers typically use the LAN only at night when numerous file transfers are performed, the raw data are archived to optical disks, and the spectral database is updated. During the day the computers are operated independently due to the large memory requirements of the LAN software. To orchestrate this rather complex data manage-

Received 6 February 1990.

* Author to whom correspondence should be sent.

† Lab Calc, a product of Galactic Industries, Salem, NH, is a PC-based spectroscopy environment which includes graphics, an array processing programming language, object-oriented plotting, and an optional searchable database module called Search. Lab Calc is used throughout our lab as a common environment for viewing, manipulation, numerical analysis, and plotting of spectra from multiple spectrophotometers.

0003-7028/90/4407-1221$2.00/0

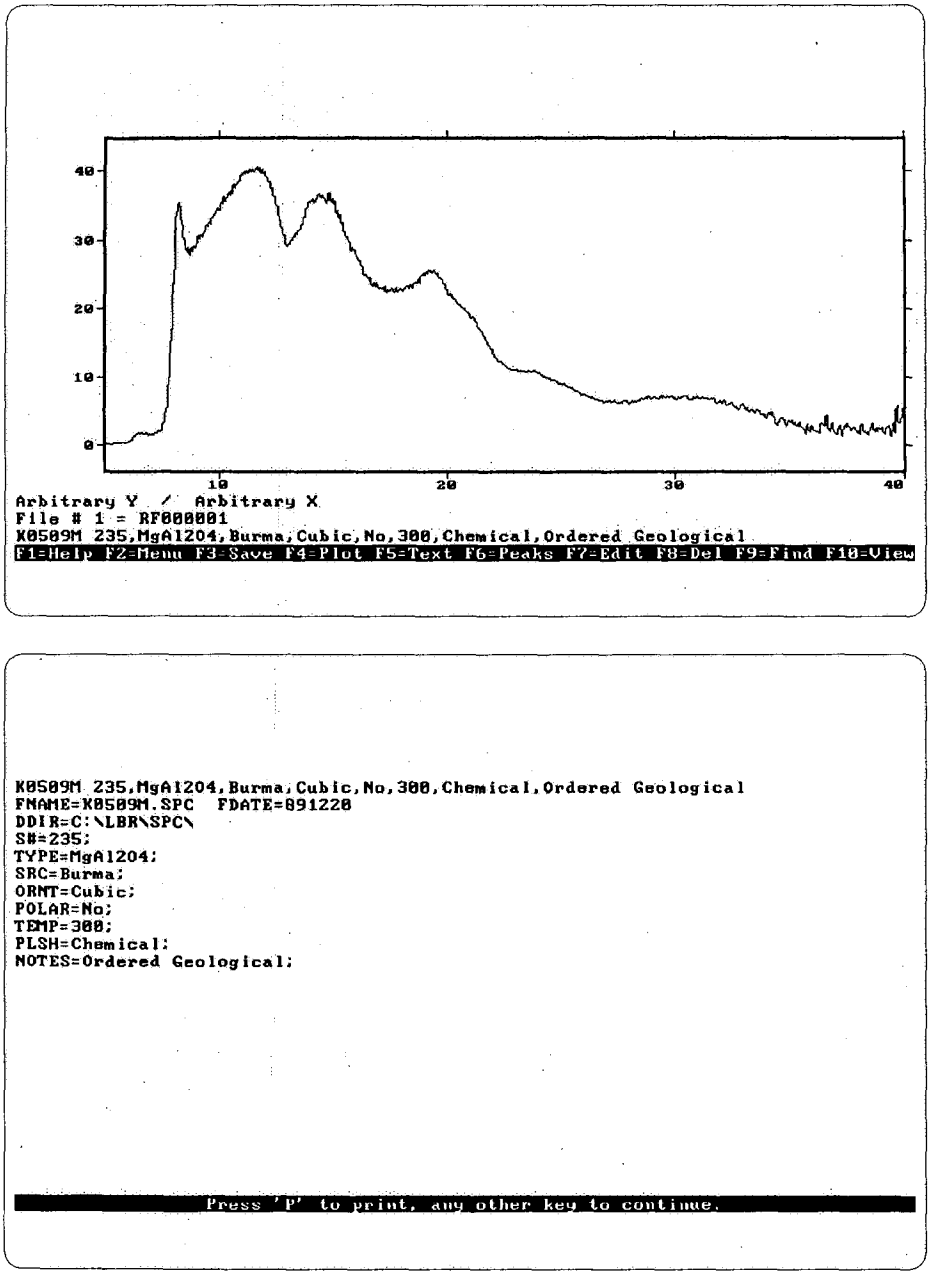

FIG 1. A typical spectral database entry created by the automatic spectral database and archive system. The spectral page shows the room-temperature optical conductivity of $MgAl_2O_4$ from 5 to 40 eV, while the database text page associated with this spectrum contains field-searchable information on this spectrum.

ment task involves the combination of a descriptive file-naming convention, encoded file header comment lines, a system-wide backup routine, and the random access file-handling abilities of Lab Calc's Array Basic programming language to create an automated spectral database.

File Name Convention. The basic organization of our research data is the bound laboratory notebook, and the computer data files which are created are named to reflect this fundamental record-keeping technique. Since the computer operating system is DOS, a file name is limited to eight characters and an extension. Lab Calc recognizes an ".spc" file name extension for data files; therefore, a spectral file is uniquely identified with the remaining eight characters. The first two identify an alpha-numeric notebook code, the next three are the notebook page, the sixth digit is an experiment number on that page, while the last digits of a file name denote the spectral data type such as dark current, lamp or sample signal, reflectance, transmission, index of refraction, dielectric constant, or optical conductivity. This convention leads to files which, in an alphabetized directory (as can be achieved with many disk optimizer programs), are logically grouped.

The file-naming convention also permits the automation of many routine analysis programs in which a pattern such as dark, lamp, or sample is required for the calculation of reflectance, or a series of experiments on one page will be concatenated. Due to this convention, an analysis program is able to suggest the next logical

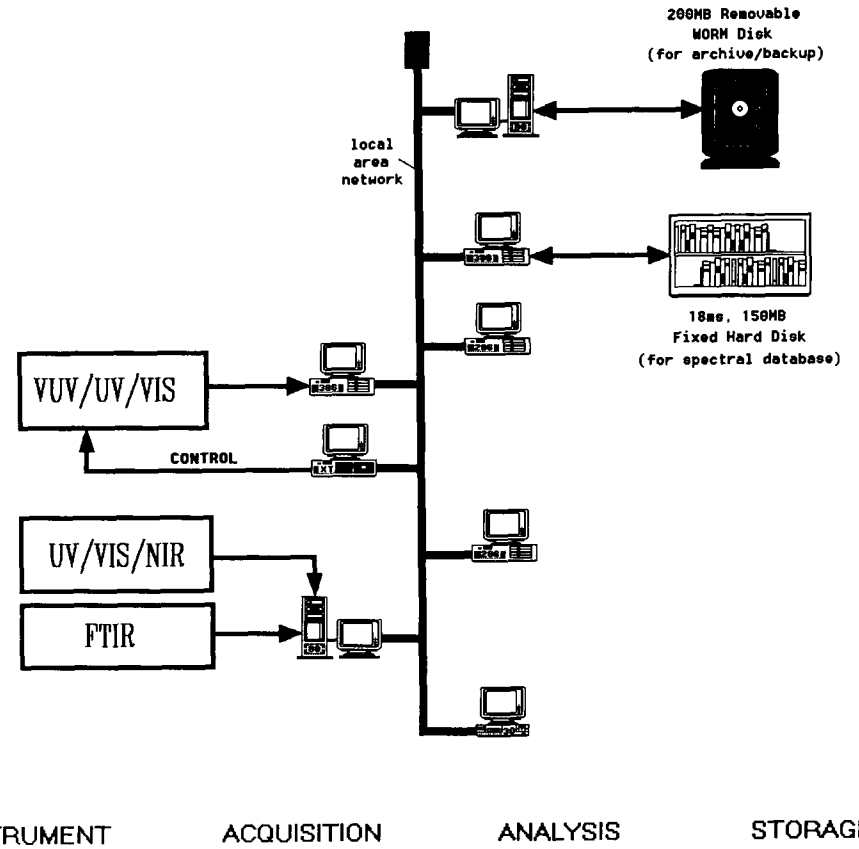

FIG. 2. Hardware layout and network topology for the automatic spectral database and archive system, consisting of the three spectrophotometers, acquisition and analysis computers, and storage devices for archive system (removable WORM disk) and spectral database (fixed hard disk).

file name which might be expected, removing the majority of error-prone file naming on the part of the operator.

Comment Lines. When a spectrum within Lab Calc is being viewed, one line of the display located near the bottom shows comment text which is stored in the header of the (.spc) spectral file. The original raw data files created by the independent spectrophotometers all support the creation of a comment line at the time of data acquisition. In our case the smallest comment line is limited to 38 characters. We use a standarized comment string format, containing sample number, type, source, orientation, polarization, temperature, polish, and notes, which is passed during file import into the Lab Calc .spc file. This comment line is displayed for any file being viewed in Lab Calc and is transferred to any files which are the result of subsequent numerical analysis. Comment line transfer is performed automatically as part of any analysis program. The comment line also provides the key word information used for building the text screen that is associated with each spectrum introduced into the spectral database, and it is the basis for rapid text-based searches of all spectral results.

Automated Backup Routine. With the various spectrophotometers used in our lab, the large quantity of data acquired, and the time-consuming nature of much of the data analysis, it quickly becomes necessary to utilize various computers and a LAN to move spectra among machines. Essential to the integrity of this system is a daily networked file backup procedure outlined in Fig. 3, the Backdisk.bat flow chart. This serves two purposes: (1) all spectra on the network, whether raw data or resultant calculated spectra, are archived on a large-capacity optical WORM drive so that any spectra can be quickly located among the various possible computers, and (2) at the same time, backups of all of the hard drives' software and directory structure are performed (not including spectra) in the event of a hard disk failure. Each evening the network is loaded, and this DOS-based backup routine runs unattended. During this procedure all new resultant .spc files (such as reflectance, transmission etc.) existing anywhere on the network are corralled into a single directory from which, upon completion of the system-wide backup, the Lab Calc Search-based automated spectral database program is launched.

Backdisk.bat copies every file which did not exist before a previous backup. This is done with file attributes and, in particular, the archive bit. The DOS Xcopy command can read the archive bit and decide whether or not to copy. Each time a file is either created or altered, the archive file attribute is set to the state which allows copying with the archive feature of Xcopy. If a file is copied with this feature, then the archive bit is toggled to prevent a subsequent copy. This method provides a means to incrementally back up the entire network in a time-saving, economical manner.

The network software optionally provides peer-to-peer access of files. (The network is not dependent on a single file server. This feature allows each computer to maintain high-speed local storage.) Before a backup begins, each network computer is placed in file server mode and shares its entire disk to the network. Backdisk ex-

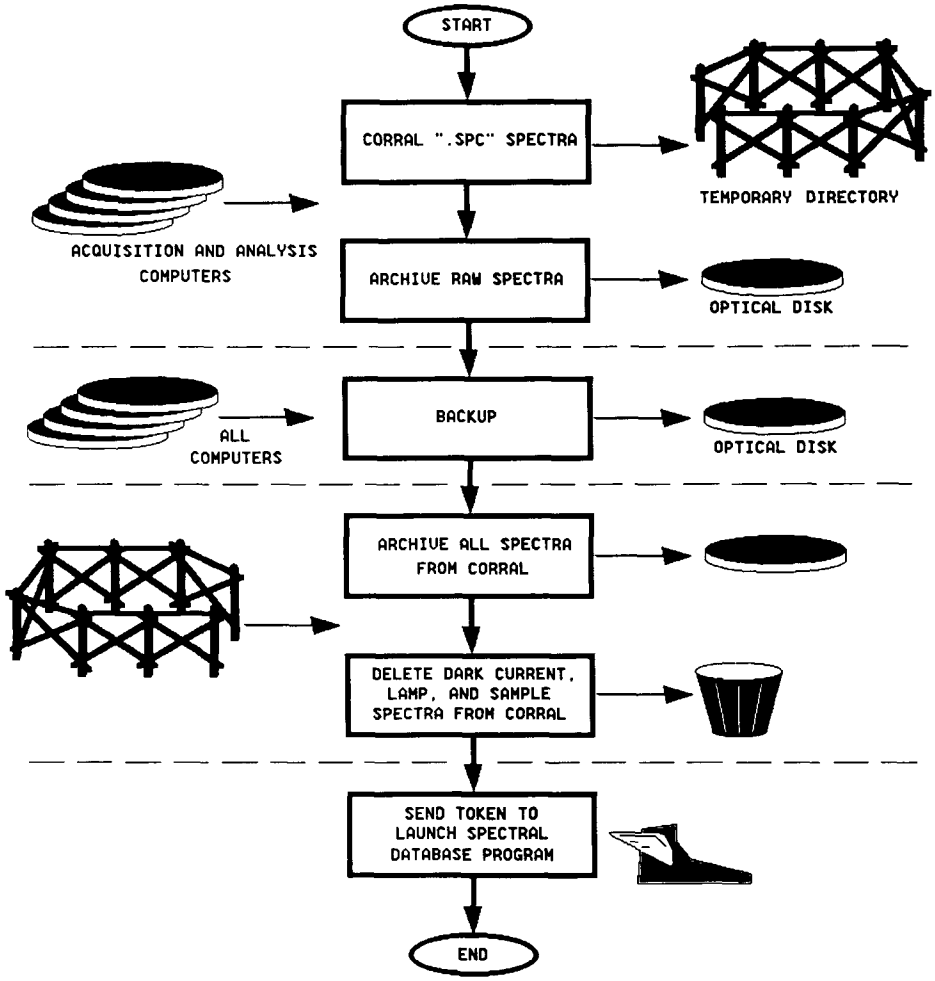

FIG. 3. Flow chart for the archive program, Backdisk.bat. Initially, spectra are corralled in a temporary directory for introduction to the spectral database, then spectra are archived. The fixed hard disks of all computers are then backed up, and a token is sent to launch the automatic spectra database program, Autolib.ab.

ecutes from the node which has the optical drive installed. The file directory structure of the optical disk is organized so that it not only reflects the organization of the network but also contains an archive directory for spectral files. This way any network node can be easily restored and spectra can be conveniently stored in one location. Consequently, the first step in Backdisk is a straight copy which archives all spectral files to the optical disk and copies .spc files to a temporary directory on a high-speed hard disk. The temporary directory is used by the spectral database routine discussed later. The archive bit is left untouched during this phase of the program. The next step is to interrogate each network disk for new files and copy them with the use of the Xcopy method. This phase is the actual backup. Finally, a token file is written to the computer that manages the database.

Automated Spectral Databasing. Certain challenges had to be faced in order to develop a program which automates the process of updating the spectral archive libraries. These include the tasks of selecting a library, parsing information from the comment line and introducing it into the database template, and handling a spectral library extraction in which a spectrum is removed from the library for futher analysis.

Each library represents the work of a single notebook. The base file name given to a library is the alpha-numeric notebook code. Because concurrent research studies are active, more than one notebook is open. Therefore, the ability to choose a library based on the file name's notebook reference convention is necessary for automation. The default selection of a library is found in the Lab Calc Search parameter file.

For inclusion of user-defined textual information in a library, a separate database template record is appended to each library entry. The information which goes into this template conforms to a file structure which consists of various predefined data fields and aids the subsequent database searches. If one wishes to enter new data in these fields, Lab Calc's interactive editor must be used or, for an automated procedure, a new template file containing the default data fields with the parsed comment line information must be generated.

The Array Basic file handling commands of Lab Calc, (i.e., open, seek, read, write, and close) have been exploited to allow one to select a library file, write a template file, and modify the spectral file comment line so that automatic archive library updating is accomplished.

Included with the Lab Calc Search package is an example program written in Array Basic, Buildlib.ab, which

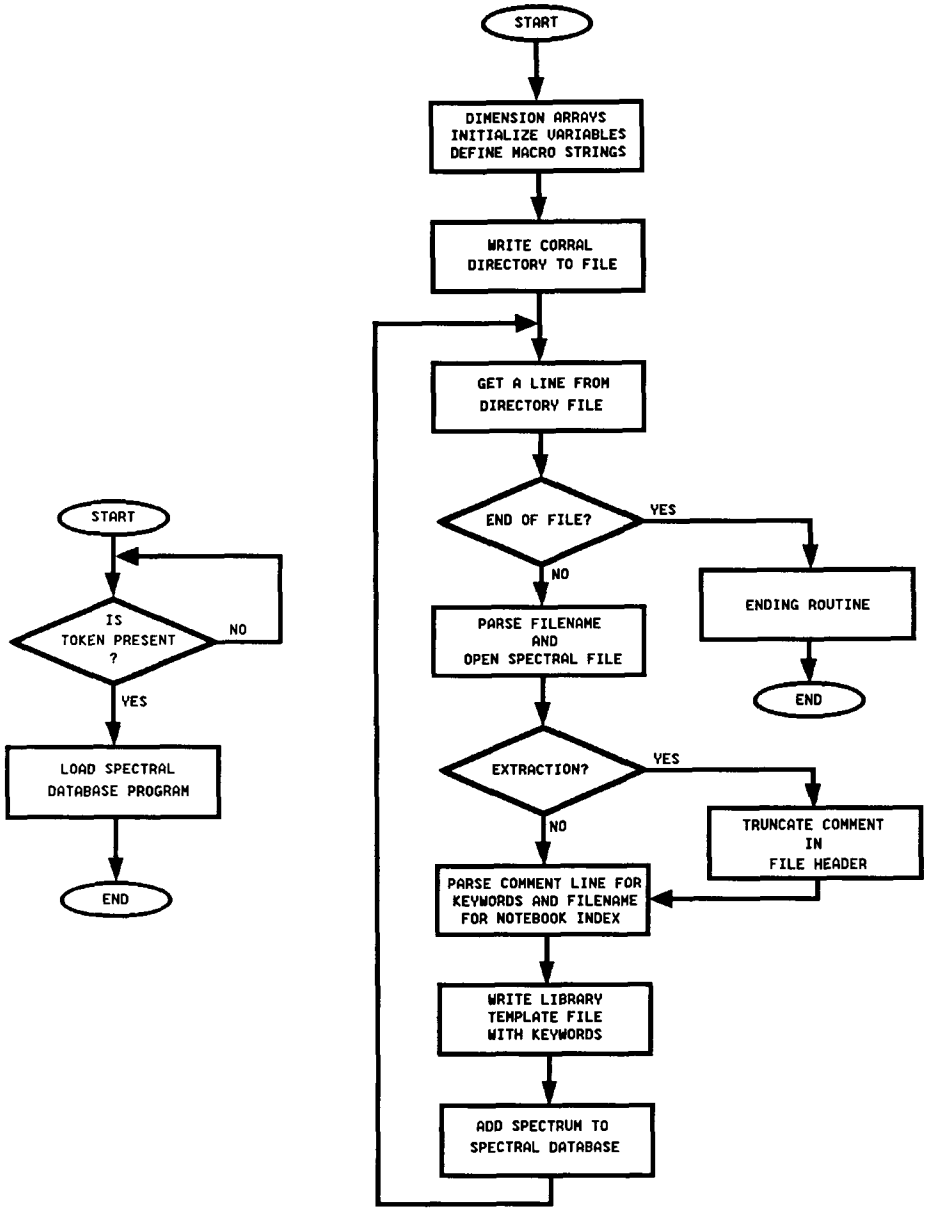

FIG. 4. Flow chart for the automatic spectral database program, Autolib.ab. Initially during archiving, the database program loops, waiting for a token to initiate databasing. Autolib.ab processes one new spectral file at a time, parsing the file name and comment line into the associated library and text page for each database entry. The spectrum is then entered into the database and the process continues to completion.

automatically enters spectra from a single directory to a preselected spectral library and appends a default set of text field information including file name and date of entry. Autolib.ab is based on this program and duplicates certain segments of its code. In addition though, the automated spectral databasing program adds a spectrum to the appropriate library for the particular research notebook and appends a searchable database template to each library entry with the default fields, each containing the parsed information from the encoded spectral file header comment line, plus any additional information introduced in subsequent analysis.

In Fig. 4, a simple DOS batch program is flow-charted on the left. Its purpose is to wait until a token file is copied from the WORM drive node to the spectral database node of the LAN. On the right of Fig. 4 is the Autolib.ab flow chart. Initially, a list of spectral files put in the corral as a result of the automated backup routine is created by writing the directory information of the predefined spectra directory to a file. At each pass of the main loop, a new line from the list is read and interrogated for the name of the spectral file to be introduced into the library. The file name is then used to open the binary spectral file for direct reading of the comment line information in the header.

The spectral file name is decoded of its notebook index and then the Lab Calc Search library parameter file is modified to include the corresponding library file. The parameter file is opened, the pointer is moved to the correct location where the path name for the library begins, and the new library DOS path name is written.

The comment line is delimited by commas and is parsed and expanded into the appropriate field information. A template file is opened that corresponds to the correct

library file, and the keyword information is written. The comment line is properly updated with a current file name, and the actual library entry is ready to occur. A macro executes a series of menu selections, and the spectrum is entered into the library. After the last file in the list is read, the program ends.

One exception to the handling of library entries is the case of a spectral file that has been created as a result of a library extraction. The comment line of an extracted file contains all of the textual data which was created for the purpose of the spectral database, all of which becomes redundant upon a possible reentry into the database. It is important to manage this unique comment line so that a subsequent entry of an extracted spectrum retains information by which one can track its history, and yet does not confuse the database structure. This is done by modifying the comment line stored in the binary file header of an extracted (.spc) spectral file. Consequently, Autolib determines whether or not the file is an extraction by evaluating the comment line. If it is, the new file name is inserted (thus, preserving the previous file name for tracking history) in the first line by driving the interactive editor with a macro.

CONCLUSIONS

We have developed an automated, systematic, spectral database and archive system based on a file name convention, comment lines introduced at the time of data acquisition, an archive program, and a database program which operates on a network of seven computers. This system handles spectra acquired from three spectrophotometers spanning the FT-IR to the VUV regions. The system supplies us with spectral archiving, hard disk backup, and a fast, text-searchable database which includes both text information and full-resolution spectral data. This automatic spectral database and archive system has proven to be a vast improvement over conventional spectral data storage and retrieval, providing an elegant solution to a complex problem. This assures us that our copious amounts of data are being preserved with integrity and supplies us with an organized structure to easily access both the data and the associated experimental information. This system also permits the automation of many file-name aspects of programs used in subsequent spectral analyses.

ACKNOWLEDGMENTS

The authors would like to acknowledge the assistance of D. S. Withrow and J. D. Simmons with the local area network, D. J. Jones for VUV/UV/Vis spectroscopy, M. K. Crawford for FT-IR spectroscopy, and D. E. Abrams and D. Kuehl of Galactic Industries.

1. M. L. Bortz and R. H. French, Appl. Phys. Lett. **55,** 1955 (1989).
2. R. H. French, "Laser-Plasma Sourced, Temperature Dependent VUV Spectrophotometer Using Dispersive Analysis," in Proceedings of the Ninth Int. Vacuum Ultraviolet Radiation Physics Conference, Hawaii, to be published in Physica Scripta **41,** 404 (1990).
3. Perkin-Elmer Corporation, Norwalk, Connecticut.
4. Bruker Instruments, Billerica, Massachusetts.
5. M. L. Bortz and R. H. French, Appl. Spectrosc. **43,** 1498 (1989).