INSTITUTE OF PHYSICS PUBLISHING MEASUREMENT SCIENCE AND TECHNOLOGY

Meas. Sci. Technol. **16** (2005) 261–269 doi:10.1088/0957-0233/16/1/034

An informatics infrastructure for combinatorial and high-throughput materials research built on open source code*

Wenhua Zhang, Michael J Fasolka, Alamgir Karim and Eric J Amis

Polymers Division, National Institute of Standards and Technology, Gaithersburg, MD 20899-8542, USA

E-mail: mfasolka@nist.gov

Received 30 April 2004, in final form 18 August 2004
Published 16 December 2004
Online at stacks.iop.org/MST/16/261

Abstract

Laboratory research informatics systems (LRIS) hold great promise in streamlining research generally, and are absolutely necessary for new data-intensive research strategies such as combinatorial and high-throughput (C&HT) approaches. In this paper, we describe a LRIS geared towards C&HT materials research, which is being established in lab facilities at the NIST Combinatorial Method Center (NCMC). The NCMC LRIS system aims to seamlessly integrate the library design, fabrication, measurement and analysis/data-mining aspects of the combi approach into a functional whole geared towards accelerating materials research. Our LRIS system is built upon an open source database management system, and incorporates non-proprietary user/instrument interface and automation software. The NCMC LRIS project endeavours to provide working examples of informatics infrastructure development to those interested in assimilating such tools into their research programmes. Accordingly, this report aims at providing an experiential account of the process by which we designed and implemented the NCMC LRIS.

Keywords: informatics, laboratory database, combinatorial, high-throughput, open source, materials research

1. Introduction

The discovery, development and optimization of advanced materials involve the exploration of huge, multi-dimensional parameter spaces and the complex correlation of materials structure (molecular to bulk), materials properties (mechanical, chemical and opto-electronic) and processing factors. To position themselves better in this complex arena, many materials and chemistry researchers have turned to the combinatorial and high-throughput (C&HT) approaches, which have recently revolutionized the pharmaceutical and catalysis industries [1]. C&HT experimentation marries the fabrication of complex multivariate specimen libraries with automated measurement and analysis ('screening') tools to accelerate the pace of product discovery and knowledge generation from complex systems. Such strategies rely on a robust *laboratory research informatics system* (LRIS) which coordinates these activities into an efficient, rapid workflow. The 'nervous system' of a C&HT lab, an LRIS integrates an experiment-oriented database, automated fabrication/characterization instruments and software tools capable of performing analysis and data-mining functions on large data sets. The promise of an LRIS is to provide a streamlined, seamless, intelligent automation that

* Official contribution of the National Institute of Standards and Technology; not subject to copyright in the United States.

incorporates and enhances all aspects of the research process. Designed correctly, such systems hold promise for increasing productivity, and enable scientists to concentrate on the more interesting aspects of laboratory research [2, 3]. Recent years have seen the development of LRIS systems geared for medical research and drug discovery [4–6], but the information technology developed for these aims cannot be directly transferred to C&HT materials research. Inherently multi-disciplinary research in e.g., nanomaterials, and advanced polymers, coatings and composites, involve complex fabrication routes and measurements using a myriad of both custom and commercial computer-driven instruments and software tools. Accordingly, LRIS systems for the C&HT materials laboratory involve a new set of rigorous and specialized approaches [7–9]. In this paper we discuss an LRIS system developed for C&HT materials research; built to enable scientific activities at the NIST Combinatorial Methods Center. Here, we aspire not only to demonstrate features of our LRIS, but also to portray the process by which we designed the system and deployed it in our laboratories with the hope that our experience can assist others in this endeavour.

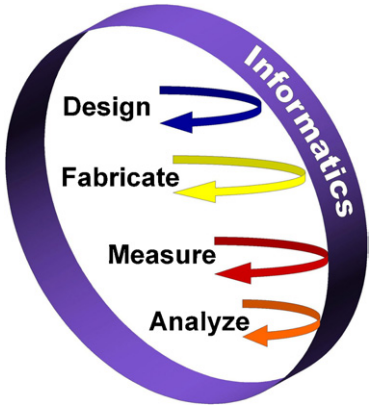

Figure 1. A combinatorial research workflow, as envisioned by NIST researchers. The NCMC LRIS facilitates this scheme by coordinating each step of the process through a centralized database. Ultimately, the LRIS will facilitate a feedback system by which high-throughput analysis of combinatorial libraries is used to generate parameters for further rounds of testing, i.e., 'closing the combinatorial loop'.

2. LRIS infrastructure design criteria

As reflected in the discussion below, implementation of a C&HT-capable LRIS is a large, complex enterprise that requires careful planning. Consideration of system design criteria sets a foundation for this planning process. In this section, we will discuss aspects of LRIS design with respect to C&HT materials research generally, and with respect to particular challenges we encountered in developing our system.

The NIST Combinatorial Methods Center (NCMC, see www.nist.gov/combi) develops library fabrication devices and high-throughput measurement tools geared towards C&HT materials research, with emphasis on advanced polymers, organic films and coatings and nanostructured materials [10–12]. As part of this mission, NCMC researchers are building a LRIS system in its combinatorial materials research laboratory facilities. The NCMC LRIS aspires to provide (1) data storage that is specifically geared to accept *materials research data* from a variety of sources, and that is *structured for scientific aims*. (2) A versatile scientific functionality produced by interfacing the central database with instrumentation and data analysis/data-mining tools. In order to effectively provide these functions, an LRIS system design must reflect the general process, or 'workflow', by which C&HT studies are undertaken. While there are several C&HT workflow models, NCMC materials research generally proceeds through four steps, which ultimately form a feedback loop by which results are used to focus and optimize later rounds of experiments (see figure 1):

(1) *Experiment design*. As opposed to traditional studies, where experiments are often conducted on single specimens in an *ad hoc* manner, C&HT research requires a formalized experiment design amenable to multivariate specimen arrays. The goals of the experiment, reflected through the library scope (see below) and the required measurement and analysis regimen, must be described via regularized parameters that can be interpreted by later automated routines. Moreover, the efficient exploration of large multivariate spaces often requires sophisticated 'design of experiment', i.e. statistical strategies [13, 14].

(2) *Specimen library fabrication*. Combinatorial specimen libraries express an array (typically planar) of compositional, structural and processing conditions spanning a large multivariate parameter space. The fabrication of combinatorial libraries necessarily depends upon automation. This includes robotic materials handling, metering, mixing and deposition devices, and instrumentation for varying processing conditions over time and/or position. Automated library fabrication must respond to experiment design parameters and must coordinate with later measurement and analysis steps.

(3) *Library measurement/characterization*. The data collection phase, this step involves hundreds, sometimes thousands, of measurements over the space of the specimen library, and often over time. For materials studies, this process often involves several types of instruments which give a variety of data types including, single numbers, spectra/profiles and images. This diversity places demands on LRIS database design.

(4) *Automated analysis*. The large amount of complex data acquired by the automated measurement of multivariate libraries demands automated analysis. This allows for usable parameters to be efficiently extracted from individual complex data types, e.g. images. Moreover, automated collation of data is essential for the visualization/correlation of trends over the library space. Indeed, in C&HT research, automated analysis is also aimed at providing parameters useful for designing later rounds of experiments. We term this feedback mechanism as 'closing the combi loop' (see figure 1).

Along with these general concerns, design of the NCMC LRIS also accommodates several challenges inherent to materials research in many laboratories, including NCMC facilities. First, in NCMC laboratories, combinatorial library fabrication most often employs continuous gradient techniques

[15], which involve specimens that gradually change in their properties/processing as a function of position. A suite of custom-made NCMC gradient fabrication devices create libraries that continuously span parameters including film thickness, composition, surface energy and temperature [16]. Automated fabrication of gradient libraries typically requires one or more computer controlled translation stages coordinated with a deposition system designed to gradually vary the rate and/or composition of deposited material.

Second, HT measurement capabilities in the NCMC involve both custom made and commercial instruments. These include automated optical [17] and scanned probe microscopy, infrared spectroscopic imaging, and contact angle measurements, and NCMC-developed screens for complex formulations properties [10], adhesion performance [11] and mechanical properties [12]. This set of instruments presents challenges regarding both the diversity of sources and data types. NCMC data sets include collections of optical and scanned probe microscopy images (variety of formats), video, two- and three-dimensional numerical arrays, and string data descriptors including units, notes and graph labels.

Moreover, in most laboratories, data analysis tools come from a variety of sources. Custom software is constructed using different languages and platforms, and commercial software most often is supplied from several different vendors. NCMC facilities are no different. Given this, integrated analysis capabilities could be achieved by reproducing the functions of each piece of existing software using a common platform. However, this strategy would necessitate a substantial (perhaps overwhelming) programming effort that would have to be renewed whenever new capabilities were added. Accordingly, the NCMC LRIS design aims to leverage existing (often sophisticated and superior) analysis routines through structures that automatically coordinate appropriate input data and that accommodate output data streams. Of course, such interchange protocols, especially ones that seamlessly integrate new components with minimal programming, are also a challenge to create.

While it is intended to streamline our own C&HT workflow, the NCMC LRIS is also meant to serve as a model or example for informatics infrastructure development. Accordingly, the NCMC LRIS is built using two principles: First and foremost, the system is constructed as often as possible with *open-source code*. Open-source software tools offer informatics building blocks that can be freely used and modified. Second, we strive to impart the system with a *transparent design*. The NCMC LRIS comprises software tools that are easily understood (in structure and function) by reasonably trained industrial or academic personnel.

In is intended that our informatics system will provide working examples on LRIS infrastructure development that can be directly used by other parties, modified to meet their needs, or due its transparent structure, compared to existing systems. This will enable feedback on the LRIS design, and illuminate issues regarding informatics standards (interchange formats etc), which the NCMC is interested in advancing [18].

3. LRIS deployment strategy

Successful construction and installation of an LRIS depends on execution of a detailed deployment plan. In our case, we chose to develop and deploy the NCMC LRIS in three stages, starting with basic infrastructure development and then adding more advanced functionality. These stages are outlined next.

3.1. Stage 1—database and database interfaces

The first stage of the NCMC LRIS deployment plan consisted of two processes

The first task was to design and create a centralized database system that is structured for C&HT materials research and that is amenable to accepting and properly organizing the diverse data sets generated from the various laboratory instruments and software routines. In many ways, a properly structured database is the most important aspect of an LRIS. Since it forms the core and foundation of the larger system, the database structure delineates the limitations of the LRIS function and its flexibility with respect to capability expansion. Moreover, the database structure ultimately defines the template by which diverse data are standardized and organized; a key issue in developing advanced LRIS functionality (e.g. inter-instrument communication). The design of an effectively structured database is not trivial. The database structure needs to accommodate data from both the current set of instrumentation, and be general enough to 'anticipate' data from instruments added to the system later. Otherwise, major components of the LRIS (especially with respect to instruments that depend on data from other instruments) will have to be redesigned each time an instrument is added. As discussed below, we attempt to impart this flexibility into our database system by designing a data structure (and data flow) that reflects the common workflow of C&HT experimentation.

The second stage 1 task involved the construction of interface structures that connect each instrument to the database, and provide graphical ports for user interaction with both the instruments and the database. As illustrated in figure 2, interface structures enable

(a) users to direct each fabrication and measurement instrument,
(b) the automated collection of diverse data sets from these devices, and the automated parsing of these data for collation by the central database and,
(c) users to efficiently query the database and retrieve data using computer stations at each instrument and at their desks (where analysis software typically resides).

As will be seen below, our graphical interfaces typically combine all these functions. Of these roles, automatic data collection/collation is the most important played by the interface systems. Automated data collection smoothes the workflow by freeing scientists from the constant transfer of floppy disks, optical media etc between instruments and to data analysis software. In addition, automated data collection serves to standardize experimental procedures, since interfaces demand regular parameter sets, including auxiliary data (e.g. environmental parameters such as temperature and humidity), necessary to fully define experiments. Typically, interface development is eased through the use of a common data interchange format that bridges the transition of information from individual instruments to the central database.

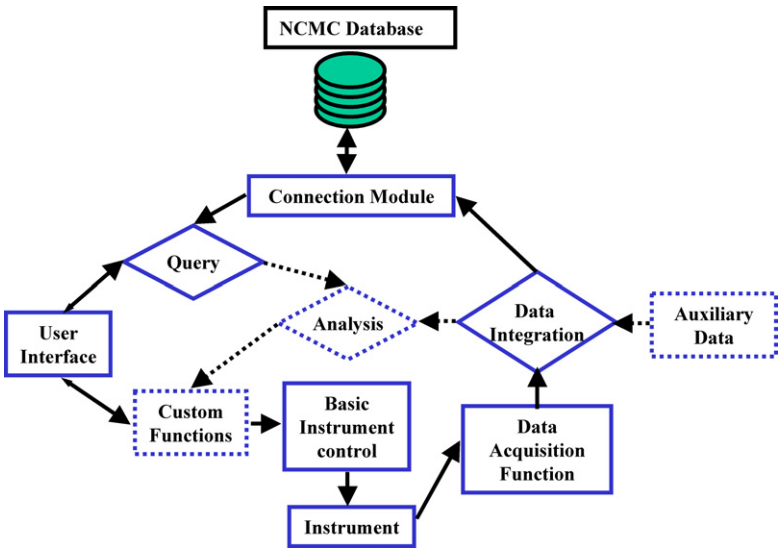

Figure 2. Flow chart illustrating aspects of stage 1 of the NCMC LRIS deployment. The chart schematically shows how connections between laboratory instruments and the central database are built. See discussion for details. (Dashed lines represent optional connections between objects.)

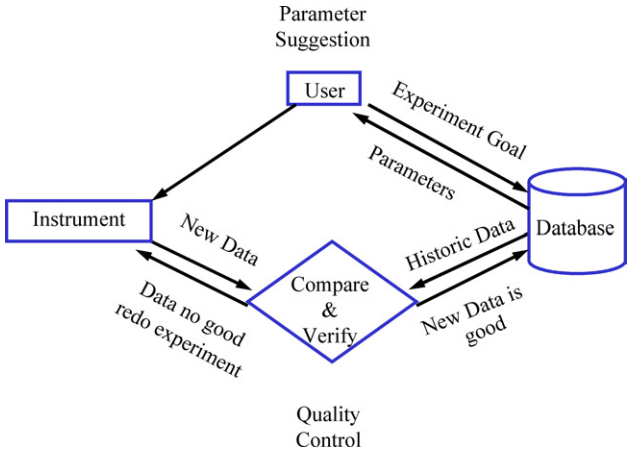

Figure 3. Flow chart illustrating stage 2 of the NCMC LRIS deployment. The accumulation of experimental parameters in the central database allows the establishment of 'expert systems' associated with each instrument. When complete, these local feedback routes aim to automatically suggest library fabrication parameters, and to serve as quality control mechanisms regarding library characterization data.

3.2. Stage 2—establishment of local 'expert systems'

Once the basic infrastructure is established, the LRIS system can be imparted with more advanced functionality. In our deployment strategy, stage 2 involves the definition of local 'expert systems' associated with each instrument. As illustrated in figure 3, our local expert systems will be designed to serve two purposes:

(a) *Suggestion of fabrication parameters.* The database automatically collects and stores input parameters from gradient library fabrication devices and data from instruments that characterize the range and slope of these continuous specimen arrays. Therefore, over time the LRIS compiles an extensive set of tables that relate fabrication parameters with library outcomes. Based on this 'experience', the LRIS database can serve as an expert system that suggests library fabrication parameters based upon desired outcomes. In this process, a user inputs the desired library design and the instrument interface queries the database to find a set of parameters likely to produce the desired outcome. For example, in our laboratory polymer film thickness gradient libraries are made through a so-called 'flow-coater' device [17]. Similar to spin casting, the range and slope of film thickness gradients is governed by a few parameters, such as polymer solution composition (polymer and solvent type) and concentration, and the motor velocity profile. Thickness libraries are characterized via a scanning interferometer, which creates a spatial map of the thickness gradient. Generating a thickness library from a new polymer often involves a few 'trial and error' cycles to achieve the desired design. However, based on previously stored flow-coater parameters and thickness library outcomes, the database can provide good first guesses for fabrication, thus saving the user time and materials.

(b) *Fabrication quality control.* The second function of an instrument-centred expert system involves quality control. This is important in library fabrication schemes that rely on automation and that are to be integrated into a workflow that aims to minimize constant user oversight. Again, compilation of library fabrication and characterization data by the LRIS database provides a growing history to which newly formed libraries can be compared. This will allow libraries with a large number of defects, or that fall outside of acceptable design tolerances to be automatically identified for subsequent closer examination by the user.

3.3. Stage 3—inter-instrument communication

The third stage of the project is to establish communication between instruments that further advances the C&HT workflow, ultimately aimed at providing automated feedback

for refined experimentation steps, i.e., 'closing the combinatorial loop'. Here, the central database is both the integrating factor of the LRIS and the mediator of inter-instrument communication, so a seamless stage 1 infrastructure is necessary. Basic stage 3 functions, i.e. automated data transfer along a few devices in a well-defined experiment chain, can be achieved through relatively simple additions to the stage 1 LRIS. Simple linear workflows demand only that data generated by a given instrument be available to the next device in the chain in a timely manner and in an amenable format. Standardized interface protocols and a properly structured database usually fulfil these requirements. Indeed, a useful LRIS can be achieved by automating a series of well-established, linear experiment chains; and this may be adequate for many moderate sized C&HT research programmes. The challenge to advanced stage 3 functionality (e.g., closing the loop) comes in defining automated routines that synthesize results towards a decision-making capability. Information technology relevant to this aim, for example the so-called genetic algorithms, has been incorporated into reported LRIS schemes [19, 20]. The NCMC LRIS does not yet include system-wide decision-making capabilities. However, stage 2 processes (which make simple decisions on the local, i.e. instrument, scale) are being developed towards this ultimate goal.

In the following sections, we provide details and examples of the NCMC LRIS structure and function at its current level of development. At the time of this report, the stage 1 infrastructure is complete. Stage 2 capabilities are being constructed. Stage 3 functionality is limited to a few linear experiment chains.

4. NCMC LRIS database system[1]

The NCMC LRIS database was constructed according to the design criteria discussed above. Our database was built using PostgreSQL [21], an object-relational database management system (DBMS) that provides adequate performance, speed, reliability and flexibility regarding extension for our system design. In choosing DBMS software for our system, we considered:

(1) Cost and portability. Since one of our aims is to provide working examples of LRIS infrastructure, we required an open source resource that we could freely use and distribute. PostgreSQL is an open source resource associated with both free web-based and commercial technical support.

(2) The ability of the DBMS to manage complex objects such as images, video and plots, which encompass the most common types of materials science data. PostgreSQL includes a 'large object' data type that treats complex data-objects as a single entry in the database.

The NCMC PostgreSQL DBMS is deployed on a Debian Linux system attached to a 1.6 terabyte hard drive array, which accommodates the large data sets (themselves consisting of potentially large objects) generated by the NCMC labs.

[1] Certain equipment, materials or software are identified in this paper to adequately specify the experimental details. Such identification does neither imply recommendation by the National Institute of Standards and Technology, nor does it imply these items are necessarily the best available for the purpose.

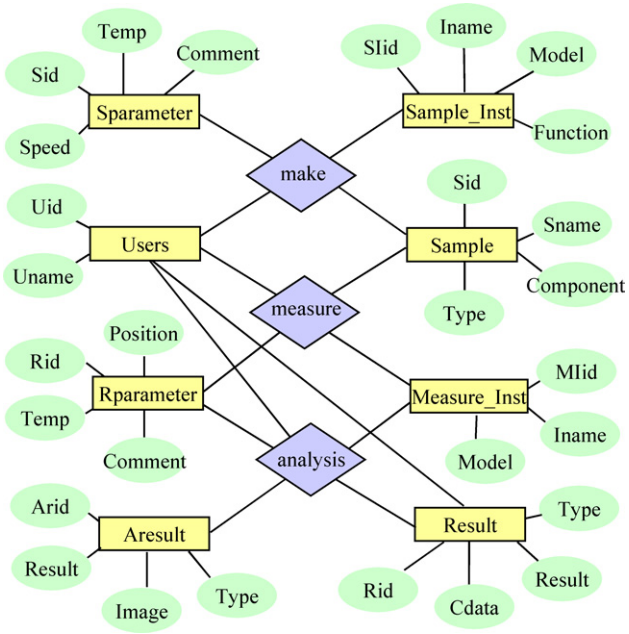

Figure 4. Entity-relation diagram for the NCMC database. By design, the database structure reflects the combinatorial workflow. In particular the centre tables, *Make, Measure* and *Analysis*, manage and organize data collected from fabrication and characterization instruments, and analysis routines. Auxiliary tables (surrounding the centre) store and transport supporting information, such as specimen identification parameters.

As designed, the LRIS database structure naturally reflects the NCMC C&HT workflow. A basic entity-relation (E-R) diagram of the database is shown in figure 4. The centre of the E-R diagram shows three database tables, *Make*, *Measure* and *Analysis*, which reflect points 1 through 4 of the C&HT procedure (see above, and figure 1). The *Make* table receives and manages information regarding experiment design and library fabrication. *Make* is linked to auxillary tables *Sparameter*, *Sample_Inst*, which organize library design parameters (thickness, temperature etc) and fabrication parameters, respectively. In addition *Make* is associated with sub-tables (*Users, Samples)*, which respectively compile user information (name, contact etc) and receive sample descriptors (sample id number, notes, components etc). The *Measure* table manages library characterization and measurement. *Measure* receives needed information from *Samples* and *Users*. An associated sub-table, *Measure_Inst*, receives measurement parameters (instrument type, model, data format etc) from characterization instruments. *Measure* provides result parameters (library coordinate of measurement [x, y], local temperature and film thickness among others) to *Rparameter*, and gives actual measurement data (library characterization) to the sub-table *Results* (see figure 4). Completing the workflow, *Analysis* receives raw data from *Results*, manages analysis routine parameters (received through an interface) and organizes processed data in *Aresults*.

Each table is designed to store the important parameters and results of the experiments in a rational way. For example, while most of the tables accept string, or numerical entities, the *Result* and *Aresult* tables have storage place for large objects including images, spectrograms and multi-dimensional

data matrices (e.g. 3D plot data). As stage 3 of the LRIS project develops, new *Design* tables will be constructed which reflect the C&HT feedback loop. *Design* will receive refined experiment design parameters derived from the *Aresults* table.

The logical similarity of the database structure and the familiar C&HT workflow makes the automation data-entry and retrieval natural to users, which smoothes operation and reduces training time. A standardized experimental procedure is enforced by logical constraints, for example to require that parameters related to the sample fabrication be sent to the database before a fabrication interface routine is closed.

As discussed above, a proper database design is essential, since it ultimately governs both the capabilities of the LRIS and the flexibility of the system with respect to growth (in particular instrument addition) and regarding access to and use of previously collected data. Accordingly, a 'piecemeal' development strategy is most often inadequate (especially for larger systems) as it involves significant programming and system redesign with each addition. The main challenge here is to design a database structure that transcends the particulars of each instrument and analysis component, and this involves taking a larger, more general view of laboratory workflow and dataflow. In our current LRIS, we attempt to meet these challenges by considering (1) common *processes* that occur in our labs and (2) common *components* of scientific data, and then reflecting these in the database structure. We have discussed point 1 at some length above. To accommodate point 2, we first considered each data type we currently or could possibly encounter in our materials research. This is a large, but limited, set that includes micrograph images, single point data and arrays of data that might be presented in 2D and 3D plots. Next, for each data type we outlined the components needed to describe the data adequately. For example, in the case of digital micrographs, one needs to include lateral dimensions (pixel/scale relationships in each dimension), pixel intensity (here we use an uncompressed tiff format), uncertainty (if possible), intensity/parameter relationship (e.g. greyscale/height multiplier in AFM images) and string data for axis and intensity units, axis and intensity labels, and notation. Similarly, in the case of arrays we considered how the data might be presented as a plot; this illuminated common components (for each dimension) such as numerical data values, uncertainty for each value, and strings for units and axis labels. With this list of data components, we were able to define a set of entries for each of our database tables and subtables (e.g. *Results*/'type'). For many of these components, the use of existing conventions can help to limit and standardize the data contained in each entry. For example, in our system string data describing units strictly conform to International System of Units (SI) nomenclature.

With this more generalized database, programming work associated with system growth is minimized. Instead of restructuring large sections of the database (or adding whole new structures), acquisition of data from new instruments (or data retrieval by analysis routines) is mediated through interface protocols written for each functional component. New functionality is thus enabled through an incremental programming effort aimed at developing a proper interface. In the following section, we will describe our strategies for writing interface programs for the NCMC LRIS.

5. User and instrument interface programming

In the NCMC LRIS, interfaces reside on instrument workstations in the laboratory and on office computers. As discussed above, these interfaces connect fabrication, measurement and analysis devices to the database, and typically fulfil functions of instrument control, automated data collection and database query and data retrieval. As demonstrated below, a given interface often will combine all these functions.

The NCMC LRIS employs the open source Python programming language [22] for generating most user interfaces, instrument interfaces and instrument automation programming. Python is well structured, relatively easy to learn and employs script editing functions that facilitate the rapid prototyping and testing of interface applications.

Python includes several specialized extension modules useful for building LRIS functions. In the NCMC system one of these modules, WxPython, was used to construct graphical user interfaces (GUIs) for Windows-based computers. Another module, pypgsql, enables connection of Python-built routines to DBMS systems accepting structured query language (SQL) queries, including PostgreSQL. Both of these modules were used to construct our query GUI (illustrated in figure 5), which allows users to selectively retrieve data stored in the NCMC database. Most users of our system have little knowledge of SQL, so our retrieval interface is designed to automatically generate SQL-compatible queries using familiar selection criteria. Via check boxes, the user selects from a series of common data descriptors (e.g. user, generation date, processing conditions, data type) that are used to generate an appropriate SQL query. The interface also allows for users to define logical conditions to better refine the range of data they query. When large objects (e.g. images) are retrieved from the database, they are either directed to proper analysis (or visualization) software or simply saved in the user's local computer. In our experience, such simplicity and flexibility reduces the training time associated with the deployment of a new interface.

As seen in figure 5, our data retrieval system allows for automatically generated SQL commands to be edited before they are sent to the DBMS. This enables users who are familiar with SQL to implement more sophisticated data retrieval and manipulation operations. As part of our development strategy, this lets us determine which SQL actions are most used (and most useful) in day to day operation—eventually many of these functions will be incorporated into the GUI. However, it should be recognized that this strategy can have considerable risks if not implemented carefully, and with security in mind. Allowing users to have direct access to database manipulation functions is dangerous, as accidental or intentional use of SQL 'delete' or 'drop' actions can destroy multiple tables in a single command. So, while we allow SQL command editing, we have installed measures that only allow the use of destructive commands by administrative users. General users have only 'read' and 'append write' access to the database, and this is reflected in the range of SQL commands they are allowed to send to the DBMS. Moreover, we protect data through daily backups of the system during off hours. As our LRIS matures, our GUI will include enough automatic functions such that

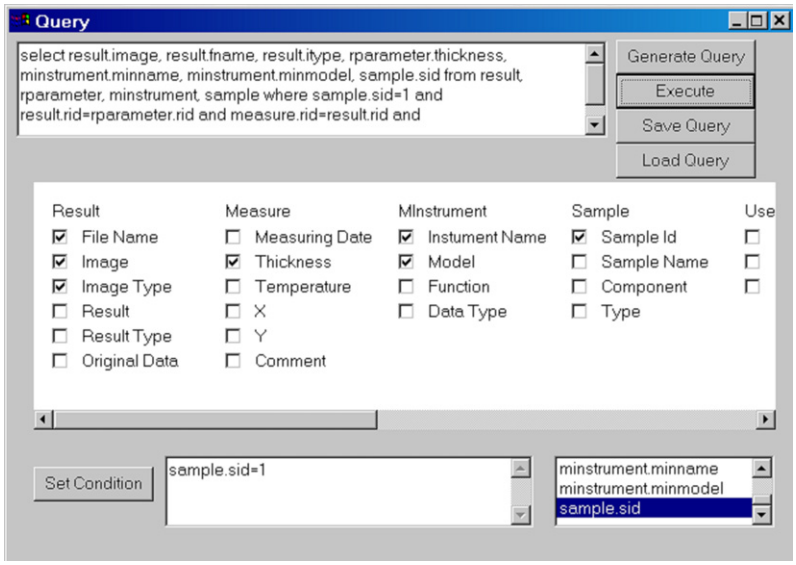

Figure 5. Screen capture image of the database query GUI. A checklist consisting of common data descriptors allows users to retrieve data without particular knowledge of SQL commands. A text input field (bottom) allows users to define parameters (e.g., ranges and logical sequences) that refine the retrieved data set. Queried data are transmitted to local computers, e.g., for batch processing by analysis software.

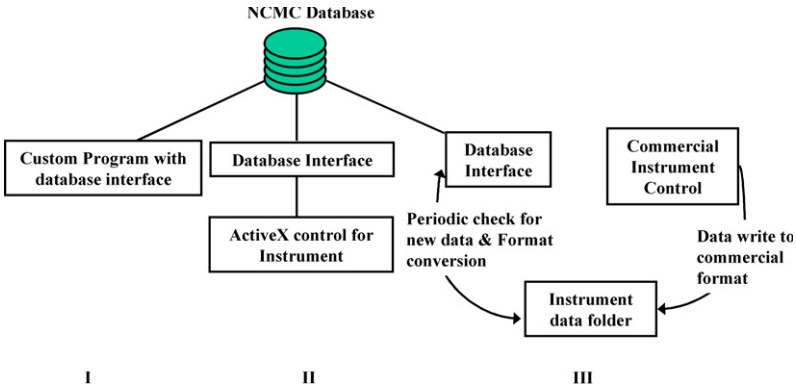

Figure 6. Schematic illustration of the three routes (I, II, III) by which NCMC instruments communicate with the database system. (I) Custom routines provide both instrument control and the database connectivity—this route is used to integrate custom-built devices into the LRIS. (II) A custom interface uses ActiveX components to drive vendor provided software packages, enabling automated control of commercial instruments and automated collection of output data by the database. (III) Here, commercial software is required to drive the device, but ActiveX routines do not exist. In this scheme, a custom interface controls the instrument through system files that manage parameter input into the device. Automated data acquisition is achieved by monitoring the software environment for new files.

direct exposure of the database to raw SQL queries/commands from users will be unnecessary and unwanted.

The development of instrument control and database attachment interfaces is often met with challenges. Custom-built devices, of course, require custom-built control and database interfaces. Few commercial instruments are supplied with software packages that have built-in database interfaces; yet the supplied software is often necessary for device operation. And while modern commercial instruments usually come with control software, it is rarely geared towards the flexible automation or coordination with other instruments necessary for integration into an LRIS. Accordingly, interface development for commercial instruments almost always involves building meta-routines that interact with and drive vendor-supplied software. Since most laboratories contain instruments from multiple vendors, each with software exhibiting different integration demands, a variety of interface programming approaches are necessary. Next, we will outline the methods we use to meet the three general interface scenarios we encountered in building our LRIS (see figure 6):

(1) *The instrument is not supplied with control and database interface software*. This is the situation for custom-built devices (e.g., our flow-coater) and for the many motor/actuator controllers. In this case, we use Python to build a custom interface, employing WxPython for the GUI, pypgsql for database connection and Python's Win32all module to manage communication through serial ports and data acquisition cards.

(2) *The instrument has commercial software and the vendor has included 'activeX' components*. ActiveX components allow Microsoft based software (see footnote 1) to be driven through external meta-routines. Many instrument vendors generate custom activeX components to enable

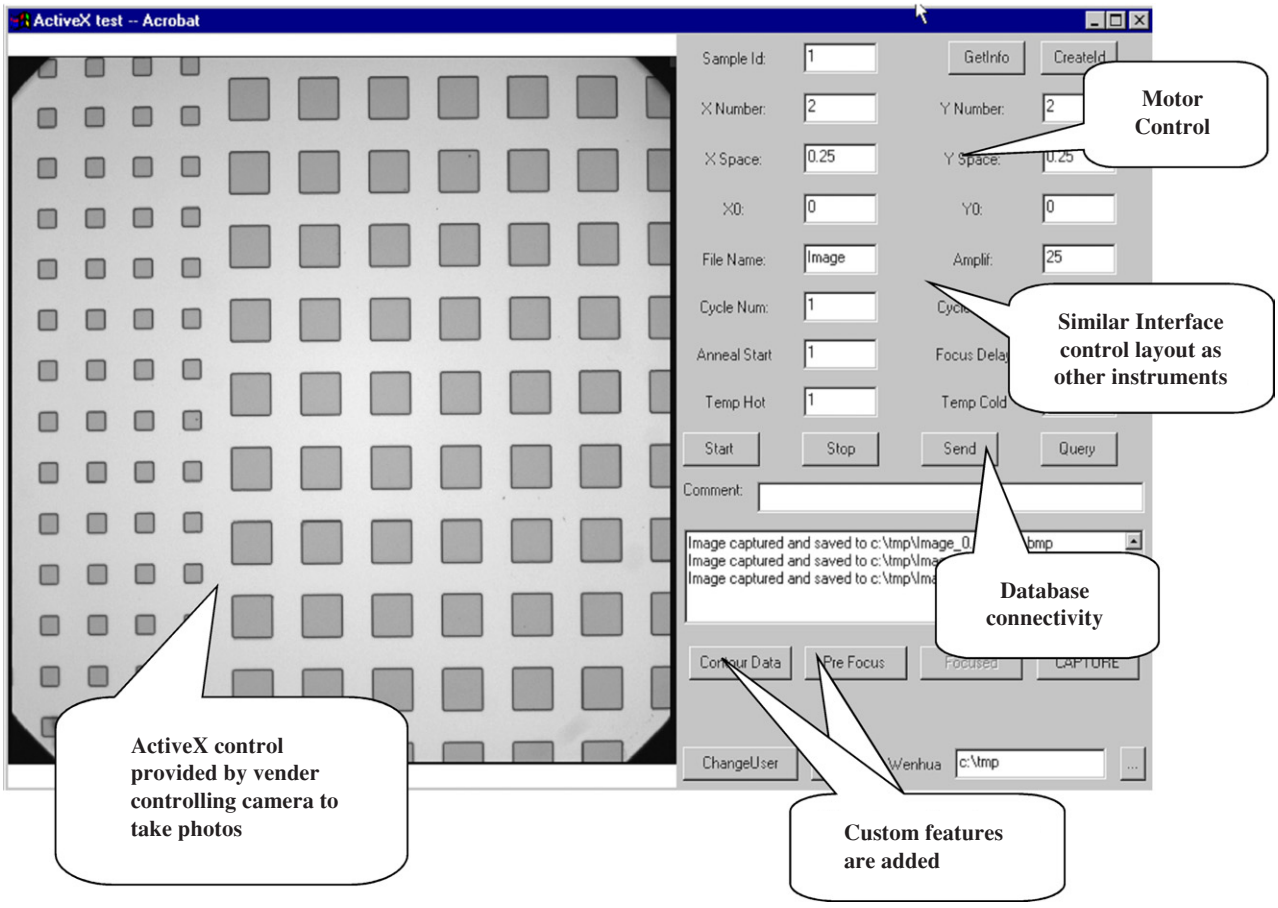

Figure 7. GUI for the NCMC automated optical microscope. The interface coordinates two activities: activeX components manage image acquisition by the camera, while custom routines drive the stage and focus motors through the serial port. The interface enables basic functions, such as image acquisition over a regular array (parameters are provided by input boxes, top). In addition, more advanced functions, such as automated focusing, are available. Database query functions (bottom) are also included.

automated operation of their equipment, but this is certainly not universal. Where such resources exist for devices in NCMC labs, a custom Python interface calls relevant activeX routines using the Python AXScript Implementation module.

(3) *The instrument requires commercial control software, but no activeX control features are available*. This is the most difficult case, and resolving it often requires investigation into the commercial software's structure and functioning mechanisms. We have found that an analysis of the file-handling functions of commercial software can illuminate routes for interface development. Here, a custom-built interface operates through the files used to input control data to the instrument and by monitoring the data output functions of the software. In NCMC labs, an example is the Python control/database interface built for our atomic force microscope (AFM), a dimension 3100 model supplied by Veeco. For instrument automation, the LRIS AFM interface leverages the parameter files Veeco uses to control micrograph acquisition; in particular the positions at which images are to be collected. In this scheme, the LRIS AFM interface processes experiment designs supplied by the user (e.g., micrograph positions) to generate appropriately formatted parameter files. These files are then inserted into the directory from which the AFM reads micrograph acquisition parameters. By providing a series of these parameter files the LRIS effectively controls the AFM by directing the instrument to acquire micrographs over defined positions in a library specimen.

The AFM is connected to the database in a similar manner. Basically, the LRIS interface constantly monitors the directory in which the AFM deposits new data files. When new AFM data are generated, they are collected and parsed by the interface, and then sent to the database in an appropriate form. During this process, specific parameters (e.g. micrograph position) are extracted from the AFM data file so the image data can be properly organized with respect to the larger library and experiment design.

Some instruments require a combination of these strategies. An example here is the automated optical microscope in NCMC facilities. Figure 7 shows the GUI for our microscope system. The microscope interface coordinates a motorized x–y translation stage, a motor-driven focus knob, image acquisition by a digital camera and communication with the central database. Here, database and motor control are achieved through custom Python routines (strategy 1), while image acquisition is mediated through activeX drivers

provided by the camera vendor (strategy 2). In addition to providing utilities geared towards basic C&HT microscopy (i.e. image acquisition over a regular spatial array), the interface combines motor and camera activities towards more advanced functions. For example, the automated microscope has a custom 'pre-focus' routine, which ensures images acquired across a library are in focus even though specimen arrays are generally not even or level. Before full microscopy analysis of the array proceeds, the interface prompts the user to provide clear focus conditions for positions across the array. The focus knob motor positions for each of these focus levels are recorded and then used as input for the array analysis. Such 'pre-focusing' is especially useful when several image acquisition cycles over the array are required as in, for example, kinetics experiments [23].

These interface structures give our LRIS basic scientific functionality by enabling direct, uncomplicated communication between instruments, users and a properly structured (i.e. generalized) central database. However, as our LRIS develops through stages 2 and 3, interface programming will become increasingly important, since the more sophisticated functionality we intend (e.g. the 'expert systems', inter-instrument communication and enhanced security discussed above) is accomplished more effectively by improving (and adding) these more flexible components rather than re-structuring the database. To achieve these aims, we are currently working towards an additional 'layer' of modular interchange components that mediate between the database and the individual instrument/user interfaces. These modules will centralize and consolidate functions that, in the current system, are completed by the individual interfaces. For example, a single 'database connectivity' module will handle data transfer between each instrument interface and the database, conversion of data to accommodate the generalized database structure and database security (access to potentially destructive database manipulations). This modular structure enhances LRIS growth, since additions and changes can be accomplished through this small set of routines, rather than reprogramming the entire set of interfaces.

6. Summary

We discussed a laboratory research informatics system being developed in facilities at the NIST Combinatorial Methods Center. The system is designed to streamline and complete a combinatorial workflow geared at accelerated materials research. Moreover, we intend this system to provide examples for parties interested in building LRISs for C&HT, so the software packages used to build the NCMC LRIS are generally open source. Indeed, we hope this paper will provide useful advice and guidance for this endeavour. To date, our progress has been in building infrastructure, in particular a centralized database system that includes connections to the instrumentation and analysis tools in our combi facilities. For more information on this system, and the NIST Combinatorial Methods Center, see http://www.nist.gov/combi.

References

[1] Schubert U S and Amis E J 2004 Combinatorial and high-throughput approaches in polymer and materials science: hype or real paradigm shift? *Macromol. Rapid Commun.* **25** 26

[2] Petersen M S *et al* 2004 Getting organized: a simple database solution to replace laboratory notebooks *Trends Immunol.* **25** 119

[3] Williams R, Bunn J, Moore R and Pool J 1998 *Report of NSF Workshop 'Interfaces to Scientific Data Archives' (May 1998)*

[4] Rabinowitz D *et al* 2004 Development of a physician attributes database as a resource for medical education, professionalism and student evaluation *Med. Teach.* **26** 160

[5] Smalheiser N R 2003 Informatics tools for the individual neurochemist *J. Neurochem.* **87** 44

[6] Panniers T L, Feuerbach R D and Soeken K L 2003 Methods in informatics: using data derived from a systematic review of health care texts to develop a concept map for use in the neonatal intensive care setting *J. Biomed. Inform.* **36** 232

[7] Yip C M 2002 Developing and implementing a unified imaging and lab information management workflow system–lessons learned and insights gained *Microsc. Microanal.* **8** 458

[8] Potyrailo R A and Amis E J 2003 *High-Throughput Analysis: A Tool for Combinatorial Materials Science* (Dordrecht: Kluwer)

[9] Rodgers J R 2003 Materials informatics: knowledge acquisition for materials design *Abstr. Pap. Am. Chem. Soc.* **226** 071

[10] Amis E J 2004 Reaching beyond discovery *Nature Mater.* **3** 83
Wu T, Mei Y, Cabral J T, Xu C and Beers K L 2004 A new synthetic method for controlled polymerization using a microfluidic system *J. Am. Chem. Soc.* **126** 9880

[11] Crosby A J, Karim A and Amis E J 2003 Combinatorial investigations of interfacial failure *J. Polym. Sci.* B–*Polym. Phys.* **41** 883
Forster A M, Stafford C M, Karim A and Amis E J 2003 Combinatorial adhesion measurements: factorial design concepts for data collection and library evaluation *J. ACS Polym. Mater.: Sci. Eng.* **88** 490

[12] Stafford C M, Harrison C, Beers K L, Karim A, Amis E J, Van Landingham M R, Kim H C, Volksen W, Miller R D and Simonyi E E 2004 A buckling-based metrology for measuring the elastic moduli of polymeric thin films *Nature Mater.* **3** 545

[13] Maxwell S E and Delaney H D 1990 *Design Experiments and Analyzing Data: a Model Comparison Perspective* (Belmont, CA: Wadsworth)

[14] Hanak J J, Gittleman J I, Pellicane J P and Bozowski S 1969 The effect of grain size on the superconducting transition temperature of the transition metals *Phys. Lett.* A **30** 201

[15] Hanak J J 1970 The multiple sample concept in materials research: synthesis, compositional analysis and testing of entire multicomponent systems *J. Mater. Sci.* **5** 964

[16] Sehgal A, Karim A, Stafford C M and Fasolka M J 2003 Techniques for combinatorial and high-throughput microscopy: part 1. Gradient specimen fabrication for polymer thin films research *Microsc. Today* **11** 26

[17] Sehgal A, Karim A, Stafford C M and Fasolka M J 2003 Techniques for combinatorial and high-throughput microscopy: part 2. Automated optical microscopy platform for thin film research *Microsc. Today* **11** 30

[18] see http://www.nist.gov//combi/CMRDS.html

[19] Hair J F, Tatham R L, Anderson R E and Black W 1998 *Multivariate Data Analysis* 5th edn (Englewood Cliffs, NJ: Prentice-Hall)

[20] Grimm L G and Yarnold P R 2000 *Reading and Understanding More Multivariate Statistics* (Washington, DC: American Psychological Association)

[21] http://www.postgresql.org

[22] http://www.python.org

[23] Beers K L, Douglas J F, Amis E J and Karim A 2003 Combinatorial measurements of crystallization growth rate and morphology in thin films of isotactic polystyrene *Langmuir* **19** 3935