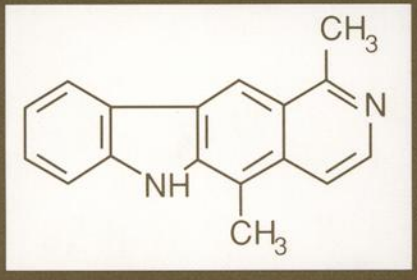
Computerized Chemical Data Standards

Databases, Data Interchange, and Information Systems



Rich Lysakowski and Charles E. Gragg, editors

ASTM STP 1214

STP 1214

Computerized Chemical Data Standards: Databases, Data Interchange, and Information Systems

Rich Lysakowski and Charles E. Gragg, Editors

ASTM Publication Code Number (PCN):
04-012140-63

ASTM
1916 Race Street
Philadelphia, PA 19103
Printed in the U.S.A.

Library of Congress Cataloging-in-Publication Data

Computerized chemical data standards: databases, data interchange, and information systems/Rich Lysakowski and Charles E. Gragg, editors.
p. cm. — (STP; 1214)
Papers presented at a symposium held in Atlanta, Ga., May 5-6, 1993.
"ASTM publication code number (PCN): 04-012140-63."
Includes bibliographical references and index.
ISBN 0-8031-1876-7
1. Chemistry—Data processing—Standards—Congresses. 2. Chemistry—Databases—Congresses. 3. Data transmission systems—Standards—Congresses. I. Lysakowski, Rich, 1957– II. Gragg, Charles E., 1951– III. Series: ASTM special technical publication; 1214.
QD8.C66 1994
025.06'54—dc20 94-31856
CIP

Photocopy Rights

Peer Review Policy

Each paper published in this volume was evaluated by three peer reviewers. The authors addressed all of the reviewers' comments to the satisfaction of both the technical editor(s) and the ASTM Committee on Publications.

To make technical information available as quickly as possible, the peer-reviewed papers in this publication were printed "camera-ready" as submitted by the authors.

The quality of the papers in this publication reflects not only the obvious efforts of the authors and the technical editor(s), but also the work of these peer reviewers. The ASTM Committee on Publications acknowledges with appreciation their dedication and contribution to time and effort on behalf of ASTM.

Printed in Scranton, PA
October 1994

Foreword

This publication, *Computerized Chemical Data Standards: Databases, Data Interchange, and Information Systems,* contains papers presented at the symposium of the same name, held in Atlanta, GA on 5–6 May 1993. The symposium was sponsored by ASTM Committee E-49 on Computerization of Material and Chemical Property Data. Rich Lysakowski of Optimize Technologies in Sudbury, MA and Charles E. Gragg of Burroughs Wellcome in Research Triangle Park, NC presided as symposium co-chairmen and are editors of the resulting publication.

Contents

Overview—R. LYSAKOWSKI vii

CHEMICAL STRUCTURE INFORMATION STANDARDS

OCR of Chemical Structure Diagrams—J. R. MCDANIEL AND J. R. BALMUTH 3

Proposed Preferred Method for Displaying Three-Dimensional Structure Information of Matter/Molecules—S. BARCZA 16

Object Data Models for Shared Molecular Structures—D. MAIER, J. B. CUSHING, D. M. HANSEN, G. D. PURVIS, III, R. A. BAIR, D. M. DEVANEY, D. F. FELLER, AND M. A. THOMPSON 32

BABEL—A Tool for Converting Between Molecular Structural Data Formats—A. V. SHAH, W. P. WALTERS, R. SHAH, AND D. P. DOLATA 45

ANALYTICAL DATA STANDARDS

NetCDF—A Defacto, Standardized Framework for Analytical Data Exchange, Storage, and Archival—R. LYSAKOWSKI 57

Evolution of an Interchange Specification for Mass Spectrometric Data—D. STRANZ, S. CAMPBELL, R. CHRISTOPHER, J. WATT, AND D. ZAKETT 75

The Demand for and the Value of a Standard, Unified Data Architecture for Analytical Testing Data—R. D. WEIMAR, JR. 84

Future ADISS Data Architecture for Materials, Properties, and Analytical Testing—R. D. WEIMAR, JR. 96

Surface Science Spectra: Practical Aspects of Applying Transformation Modeling to Analytical Data—R. N. LEE AND S. W. GAARENSTROOM 107

Considerations in the Maintenance and Dissemination of the Powder Diffraction File Database—R. JENKINS, M. M. FORNOFF, M. A. HOLOMANY, AND M. M. KING 117

DATA ACQUISITION, SEARCHING, AND REPORTING

An Information Architecture for Composites Design and the Role of Knowledge-Based Systems and Product Data Exchange Standards—J. K. MCDOWELL, M. C. HAWLEY, A. KAMEL, AND J. STICKLEN 129

The Use of SGML for Retrieval of Chemical Data—R. A. STREHLOW, T. O. TALLANT, J. D. MASON, P. L. KIENLEN, AND K. T. BARRY 138

Visual Presentation of Numerical and Scientific Data in Printed Publications—B. LAWYER AND M. KING 146

Overview

Most manufacturing industries are heavily dependent on chemistry for some part of their operations. Many formulate their own chemical products and use some kind of chemical structure modeling or database techniques in the process. To be competitive, industries must computerize as much of their chemical data as possible. Companies can actually rise above their competitors by giving the right tools for rapid, on-line information access to their most talented people. Computerization of chemical information has been a competitive strategy for many industries for many years. Yet, with this acceleration of chemistry into the information age, the creation and approval of formal standards for database storage, interchange, and presentation of chemical data has been disproportionately lagging. While standards for new technologies should lag long enough for those technologies to mature and undergo natural selection, a much stronger focus on communications standards has been needed for several years. Few formal standards now exist for exchange of chemical structure data, analytical chemistry data, or for presentation of chemical information to end users.

The first international symposium on Computerized Chemical Data Standards: Databases, Data Interchange, and Information, held in Atlanta on 5–6 May, 1993 underscores the importance and the emergence of standards for chemical data. Chemical and computer professionals from all over the world gathered to exchange the latest information on their work to establish chemical data standards and on what difficult problems are being encountered and solved along the way.

The chapters in this book are from the leaders in the field of chemical data computerization. Each have personally struggled to create and advance models and solutions to pressing chemical information problems. The job of any leader is to focus attention on the most urgent and pressing problems, to keep us abreast of recent and important developments to solve those problems, to provide a clear vision for the future that has those problems ameliorated, and most importantly, to provide a path forward to attain the vision.

In this symposium, there was a strong undercurrent of need, the need for more widespread understanding of the issues and the need for better, more comprehensive standards for computerized chemical data standards. Although this first symposium was relatively small, the papers presented were high quality and important, and in some cases, seminal. The symposium covered chemical data interchange, chemical databases, and chemical information systems. It was divided into three segments: (1) molecular structure information, (2) analytical (testing) laboratory information, and (3) presentation and publication of chemical data.

Managing Complexity with Standards and Abstract Models

The types and uses of chemical data are so varied that establishing standards is a very difficult and complex task. The accelerating and relentless application of computers to the advancement of science, technology, and society magnifies this complexity, which greatly magnifies the need for standards. Without a standard "vocabulary for sharing," the ability to compare and check data grows ever more tedious and difficult. However, order can be extracted from this growing complexity by applying the right levels of abstraction and data modeling. To create models that work well, the models must first be considered as part of a larger system into which they fit.

A common theme during the symposium was that systems modeling and abstract data modeling are requirements for current and future standards in chemistry. Techniques like abstract data modeling, object-oriented analysis, design, and implementation are absolutely essential to break complex systems into modules that can be easily integrated, managed, and supported repeatably and reliability in daily operations and evolved gracefully over time. This pervasive trend toward object-oriented approaches for designing and implementing chemical data standards is deliberate and important, but for reasons that are seldom appreciated, if ever articulated.

Moving Chemical Data Standards into Modern Times

It is difficult to pick up a software publication now without seeing references to "objects" or "object-oriented systems." These terms have become so overused that their true meaning is often lost in the journalism. What object-oriented should mean to the layman is that software systems are designed to be nearly a direct reflection of objects in the real world. Physical objects in the real world have well-defined functions. Their internal complexity is normally self-contained and not exposed to the outside world, except under conditions of surgery or detailed reductionism. To the computer scientist this is termed "information hiding" or "complexity hiding." Object-oriented system components are modular, meaning they serve a limited number of functions and have well-defined interfaces to the other modules in the system in which they exist. Object-oriented systems have two other primary characteristics that differentiate them from previous software design approaches, those of "inheritance" and "polymorphism." Inheritance permits skilled designers to *gracefully* evolve system components, or even entire systems, by actively taking advantage of what has been done before. Object-oriented software compilers actively support code module reuse through inheritance. Polymorphism is the ability to select software methods at runtime, depending on the context (the exact object type) from which they are called.

You may wonder why you have this brief introduction to object-oriented systems at the beginning of a book on chemical data standards. This point has never been made strongly enough in the standards community, that object-oriented design principles and the requirements of standards represent the perfect union if ever there was one. At least four very practical reasons exist to base chemical data standards (or any standards for that matter) on object-oriented concepts and technologies.

1. Standards are sometimes obsolete even before specifications are finally agreed upon by a standards committee. This is also true of any software technology. One may take the view that any finished software product is simply the prototype for the next version of that product. Standards are the products of standards organizations; version control in standards is just as important as for any other type of product. Version control allows standards to be finished. By definition a standard is a version with boundaries that excludes features and functions users wanted. Though software is somewhat ethereal and much more easily changed than hardware products, there are still the real world boundaries of available time, money, and talent. Mechanisms must be built into chemical software standards to allow them and their implementations to be extended easily and gracefully.
2. Another problem with the way standards have been implemented in the past is that vendors' extensions to the standards end up making them non-standard, that is, they cease to allow products that "conform" to the standard and continue to interoperate. A mechanism is needed where vendors can use these software standards in their products, apply them to fit the specific needs of their customers and give them flexibility to build in competitive advantages, yet still allow them to interoperate with other products based on the standard.

For analytical instrument data handling software, the interfaces to data communications and storage systems used in vendors' products must conform to specifications for a standard "data communications and storage standard module." The standard interface preserves the function of the standard to allow the product to interoperate with software packages from other vendors. And, by having vendors all use object-oriented technology to implement the core standard communications and storage modules, vendors can make their extensions as needed, again without interfering with the functioning of the standard messages or standard interfaces, since their products inherit all the methods and data structures provided by the interchange and storage objects.

3. Chemists in particular have a unique problem with the way that they design analytical instruments. Instrument designers take "old" instrument types such as chromatographic or mass spectrometric instruments, and combine them in new instruments, typically called "hyphenated instruments." Combining instruments in this way, whereby the new instrument inherits the characteristics of the old instruments, suggests a definite kind of inheritance that the chemical instrument industry has not exploited yet. Data systems for these hyphenated instruments inherit nearly all of the data structures and data access methods of the old instruments, and add new ones to handle the physical sample interface and data interfaces between the two instruments. Additional data structures and methods are needed to control and take data from the sample interface itself, but the bulk of the data and methods for operating on the data are the same. The need for inheritance here in design and construction of software should be obvious, yet why aren't more manufacturers taking advantage of this approach in their software businesses? Similar problems will be encountered in "computational instruments" being developed to simulate experiments before costly chemical synthesis and testing occurs.

4. Chemical instrument systems exhibit a large degree of semantic "polymorphism," that is, the same term can mean many different things, depending on the context of term. Overloading of meaning occurs because of omnipresent limitations in human language.

Although English is one of the richest languages of the world because it draws on so many other languages, it still has a limited number of English language terms for related niches of science. This is both good and bad for users of these terms. Polymorphism in a software language is the concept of having a single operator or software method execute differently depending on what the exact instance (or instantiation) of software module or abstract class.

For example, for an abstract software class called a virtual chromatograph, there may be certain commands or methods defined, such as, "prepare separation column for sample injection" or "run a sample." In the instance of a gas chromatography (GC) column, this would mean ramping up the column to the proper injection temperature. In the instance of a liquid chromatography (LC) column, this would mean ramp up the solvent gradient to the proper ratio for a particular sample. In both cases, the command is identical, but very different instantiations (instances) of the command execute, each of which depend on the context in which the command is executed. Polymorphism in software is another object-oriented systems concept that allows a common core of code to be reused for similar classes of business problems.

These four problems suggest technical solutions for standardizing data interchange and storage portions of software applications for chemistry. The three concepts of modularity, inheritance, and polymorphism produce graceful evolution for chemical data standards and software based on them. The main design premise of object-oriented software systems has been for the past 20 plus years, graceful evolution of software to meet evolving needs of users at all levels, from end users, to application programmers, to systems designers. The design attributes of object-oriented systems fit the characteristics of well-defined standards better than previous approaches to standards implementations. In fact, this is a better marriage of

technology and standards than we have ever seen in the past. The chapters in this volume detail some of the first applications of these concepts to chemical data standards.

Brief Overview of Chapters

The chapter by J. McDaniel and J. R. Balmuth of Fein-Marquart Associates, entitled *OCR of Chemical Structure Diagrams* describes a concept implemented in the Kekule product that promises to be a great time-saver for chemical professionals doing computer entry of chemical structures from published literature or hand-drawn structures. This chapter describes the concepts, strengths, and weaknesses of the technique, whereby any scanned image can be transformed from a bit-mapped image into a computer data structure maintaining full chemical significance.

Proposed Preferred Method for Displaying Three-Dimensional Structure Information of Matter/Molecules by S. Barcza of Sandoz Pharmaceuticals presents a comparison and review of the methods for 3-D rendering of chemical structures to the human eye. Dr. Barcza posits an inexpensive method for publishing structures on paper that can then be viewed as 3-D structures by the unaided human eye. As you will see from the instructions and examples, the method works and is safe for occasional use. The paper is unique and refreshing in its presentation of the material, because it subverts society's rampant overemphasis on "high-tech" solutions where sometimes the "low-tech" works fine.

Object Data Models for Shared Molecular Structures by D. Maier et al. presents a striking example of the benefits that object-oriented systems modeling techniques can bring to chemists. Most automation tools used by researchers today force-fit the chemical researcher's paradigm of work into financial spreadsheets, word processors, and other office automation tools. However, this interdisciplinary software development team has modeled chemical research as a collection of chemical data objects and processes, and developed tools that closely match chemists' thinking about the process. Rather than struggling with tolls that don't quite fit chemical data and processes, chemists can now start with a common conceptual model defined in their common objects research architecture (CORA). From this model, integrated systems for chemical research project management, computational chemistry, and materials science research can be built that promise to make chemists more productive.

BABEL—A Tool for Converting Between Molecular Structural Data Formats by A.V. Shah el al. presents the results of a grand piece of work that has long been needed by the molecular modeling and computational chemistry communities. BABEL is a conversion hub approach to data communications that side-steps the need for molecular structural data standards. While such "standard" approaches as SMD, JCAMP-CS, SMILES, MOLFILE, and others jockey for position in global standards committees, the BABEL system fills an immediate need for a system that gets the job done. It provides a split converter hub approach similar to the one patented and popularized in Digital Equipment Corporation's Compound Document Architecture (CDA) converter hub, which cuts the job of converting to/from a new format in half. BABEL provides input and output converters for most of the popular chemical structure formats. All data passes through BABEL's universal molecular structure (UMS), which is the in-memory data structure for mapping between formats. Chemical structure formats not yet supported by BABEL can be accommodated by creating new converters for either input, output, or both. Scientific programmers can extend BABEL's UMS if it lacks certain chemical structural characteristics, because full source code and documentation are available from the authors. Everything in BABEL is public-domain and contributions to the code and knowledge base are encouraged. The design, implementation, and dispersal of BABEL has been a truly collaborative effort among computational chemists and modelers, arising from a real need for free conversion tools to make chemical structural data communication easy. Their efforts should be lauded with widespread usage.

NetCDF—A Defacto, Standardized Framework for Analytical Data Exchange, Storage, and Archival by R. Lysakowski of Optimize Technologies discusses the instrument industry's adoption of a public-domain software system for interchange, storage, and archival of scientific data, called netCDF. NetCDF was developed by the Unidata Program Center in Boulder, Colorado for large, multidimensional meteorological datasets. Scientists use netCDF for many other applications, including oceanography, space science, CAD/CAM drawings, medical X-ray imaging, and as a general-purpose storage mechanism for visualization of large scientific datasets. The author recommended netCDF to instrument industry leaders, along with a generic analytical information system model, called the analytical data interchange and storage standards (ADISS) architecture. ADISS/netCDF has now been adopted and implemented by nearly all major analytical instrument companies. In this paper, you can read about how netCDF fits into the ADISS architecture and many other reasons why netCDF has become a defacto standard for chromatography, mass spectrometry, infrared and UV-visible spectroscopy, and soon will be used for several other routine analytical chemistry instrumental techniques. This chapter illustrates the fruits of direct and efficient technology transfer between government, universities, and industry over Internet.

Evolution of An Interchange Specification for Mass Spectrometric Data by D. Stranz et al. discusses the current mass spectrometry instrument manufacturers' data standard and how it developed from an earlier protocol. The American Society for Mass Spectrometry (ASMS) developed a specification for mass spectrometry data based on a protocol, called JCAMP-DX. The JCAMP-DX protocol was eventually found to be unreliable after extensive testing because it was incompletely specified and no robust and supported software was ever made freely available to the scientific community to support JCAMP-DX. The mass spectrometry instrument manufacturers then switched over to ADISS/netCDF. Specifying the original mass spectrometry data elements was very hard work. Fitting them into the ADISS information model was relatively straightforward. Implementing data transfer required only a couple of weeks for nearly *all* of the major mass spectrometry instrument vendors. This chapter recounts the history of development of the ASMS' data standards and describes many technical details of the standard, which recently became the international Analytical Instrument Association's ANDI/ Mass Spectrometry Data Standard.

The chapter entitled, *The Demand for and the Value of a Standard, Unified Data Architecture for Analytical Testing Data* by D. Weimar, Jr. contributes to the literature because it details the high-priority business requirements and advantages of building a unified data architecture. It explodes many myths about advantages of proprietary information architectures. Although it starts in the testing laboratory with analytical data, this chapter provides many reasons for continually driving an organization toward data standardization at all levels. In the words of this author of several important data standards, "It ain't cheap, and it ain't easy, but you still need it."

Future ADISS Data Architecture for Materials, Properties, and Analytical Testing by R. D. Weimar describes the initial unified data architecture (UDA) for materials, properties, and analytical laboratory testing data, which is a larger framework for the original ADISS Architecture proposed and adapted by the Analytical Instrument Association for its standards for chromatography, mass spectrometry, and infrared spectroscopy. The author covers the principles and methodology for development of the broader ADISS data architecture, which are enterprise data modeling and entity-relationship modeling; they apply in this case not just to database models, but also to data communications. This paper discusses the detailed technical work to date in developing this UDA for ADISS beyond the analytical laboratory.

Surface Science Spectra: Practical Aspects of Applying Transformation Modeling to Analytical Data by R. Lee and S. Gaarenstroom is a seminal work on developing a standard mathematical methodology for describing analytical measurement processes, called transformation

modeling. This chapter discusses the problems that the surface scientists have had applying traditional data modeling techniques. Surface analysis instruments are perhaps the most difficult of all instruments to standardize because the instrument's physical configuration can literally change with each experiment. Transformation modeling is a way to develop a "parameterized functional object model" for each component in an instrument system, and then combine these to have an exact model of the instrument. The methodology is very general and can be applied to any sequential process whose individual steps are amenable to exact functional description. It resembles transform modeling used in control systems theory to design process control systems for chemical manufacturing environments. However, the transformation modeling approach is unique. It advocates using a mathematical formalism for describing measurement processes that could lead to a type of standardization qualitatively different from data specification; namely, requiring each instrument manufacturer specify *exact and complete* transfer functions for every measurement module that they sell. These would make it possible to calculate a system's exact performance for an application by knowing its components' transfer functions, simply by plugging all transfer functions into a system evaluation program. Incompletely or inaccurately specified transfer functions become immediately evident to the purchaser. Thus, standardization and purchasing of instruments could become exact sciences. This paper introduces a theoretical basis to make this possible.

The paper entitled *Considerations in the Maintenance and Dissemination of the Powder Diffraction File Database* by R. Jenkins et al., of the International Center for Diffraction Data (ICDD) covers the ICDD's efforts over the years to make X-ray powder diffraction data available. Initially, their database was stored and distributed in paper cards. With the advent of computerized ways of storing and distributing data came the problems of continually changing computer media standards. This paper reviews the methods and the reasoning used by the ICDD in migrating from paper cards, then films, to compact disks. The long-term maintenance of such reference databases is a big problem with which the ICDD deals very effectively.

The paper entitled, *An Information Architecture for Composites Design and the Role of Knowledge-Based Systems and Product Data Exchange Standards* by J. K. McDowell et al. of Michigan State University presents a design for building distributed polymer composites knowledge bases. The knowledge base design uses a software architectural approach to combine shared vocabularies developed using the product data exchange standards (PDES) with "domain-specific" interfaces to the knowledge base. This gives an engineer working as a "materials compositor" a way to build new materials from off-the-shelf materials specifications more quickly. The author proposes an innovative software architecture that includes problem-solving application layers, going through a domain-specific layer of software that understands the standardized materials language of PDES to access a common set of knowledge bases. The author discusses his research group's progress in testing this architectural model.

In the paper entitled, *The Use of SGML for Retrieval of Chemical Data,* R. Strehlow (Termco, Inc.) et al. propose using the structural aspects of the structured generalized markup language (SGML) for defining standards for publishing chemical and materials property data. The advantages of this approach are that any textual documents and data could be structured, indexed, and hence made fully searchable using commercially available text retrieval engines. ASCII text is the only truly universal computer data exchange mechanism in existence today; unfortunately, it provides no data structure conventions of its own. Despite millions of dollars invested by several large corporations to establish their own document standards, SGML remains the only internationally accepted formal document exchange standard. SGML provides data structuring conventions for formatted text in the form of document type declarations (DTDs). Defining these DTDs for analytical and chemical structural information for reporting scientific data in accordance with a fixed structural form would help dispel some of the confusion

around presentation and publication formats. This chapter briefly gives some background and the advantages of this approach.

Visual Presentation of Numerical and Scientific Data in Printed Publications by B. Lawyer and M. King of the International Center for Diffraction Data (ICDD) discuss their solutions to problems of publishing large amounts of numerical data directly from databases into books and reports. This type of publishing can present very difficult problems of poor readability and subsequent usability. Careful design of the layout, fonts, colors, use of white space, and other factors are essential. The ICDD's use of desktop publishing tools tightly integrated with a relational database greatly enhanced their ability to prototype and test designs before implementing database changes. It also served as a good production vehicle when design phases were completed. This chapter nicely portrays just how much of an art form it is to publish numerical scientific data with great usability.

The Future of Chemical Data Standardization

Various industry groups working on chemical data standards, such as ASTM, the American Chemical Society (through the Chemical Abstracts Service), the Analytical Instrument Association (AIA), the Laboratory Automation Standards Foundation (LASF), the Analytical Data Interchange and Storage Standards Project (ADISS), and others are achieving a respectable level of global standardization. More work needs to be done to coordinate and harmonize the results of these groups, as well as work more closely with other efforts such as the International Standards Organization's (ISO) Standards for the Exchange of Product Data (STEP), the Department of Energy's Containment Analysis Automation (CAA) Project, the United States' National Institute of Standards and Technology' Consortium on Automated Analytical Laboratory Systems (NIST CAALS), and others.

More harmonious chemical data standards are possible within the next five years. However, before that can happen, education and implementation gaps must be filled. What has been done needs to be more widely and quickly disseminated to end users. More software, documentation, tools, courses, and technical workshops can fill this hole, so that standardization can proceed more easily. The importance of the right enabling tools for developers and end users, and a political and business climate that demands open and comprehensive standards from commercial software vendors must not be underestimated. Without the confluence of all these things, standards take too long and sometimes become obsolete or die before they reach the consumer. This volume is intended to promote the educational process and make people aware of where we are today, point to ongoing developments, and point to likely areas of future development.

Richard Lysakowski
Optimize Technologies
8 Pheasant Ave.,
Sudbury, MA 01776;
symposium chairman and editor.

ISBN 0-8031-1876-7