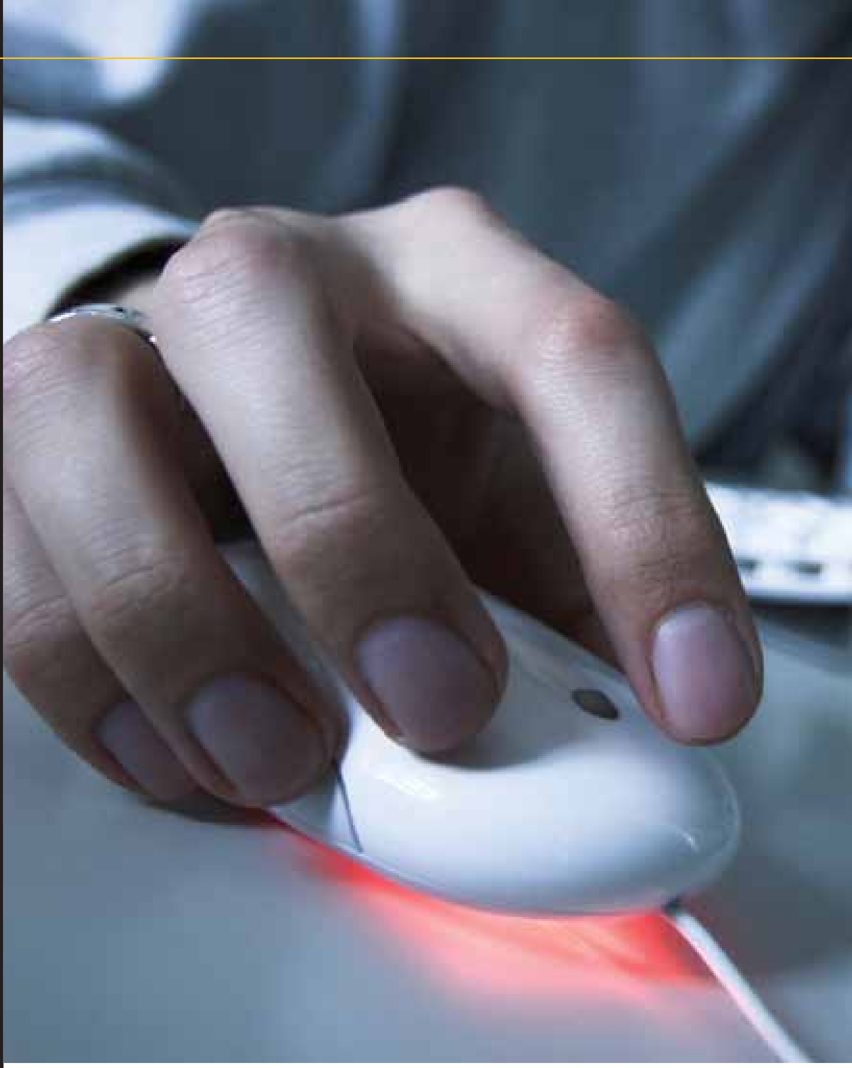
A perspective on materials databases

Stephen Freiman, Lynnette D. Madsen and John Rumble

Improvements in measurement procedures; lack of materials specifications, especially with respect to ceramics; access to proprietary data on new materials; and the need for quality evaluation are the important issues involved in developing and maintaining materials property databases.



Recently, the U.S. National Science Foundation announced that all future proposals need data management plans. Although database additions may only cover one aspect of data management, the entire issue takes on increased importance for NSF grantees.

It is widely accepted that convenient access to reliable materials property data is vital to the development of innovative components and devices. These data are necessary for the design of new and complex multimaterial structures and for the development of hitherto unknown materials, as well as for combinations of materials. Unfortunately, at this time, property data on new and emerging materials are either widely dispersed or entirely unavailable to the commercial, academic and government research communities. Data that are available are frequently unevaluated or obtained through unknown or unreliable measurement procedures. Uncertainty regarding the accuracy of such data can be an impediment to its use.

Moreover, modeling and theoretical calculations of properties are becoming more frequently used instead of expensive and time-consuming physical measurement. One of the conclusions of a recent National Research Council report on Integrated Computational Materials Engineering[1] is that databases are required for capturing, curating (culling and selecting) and archiving the critical information required for the development of the ICME area. The field of ceramics can be used as a prototype for the issues facing the development and maintenance of a materials property database in today's environment. Herein, ceramics are defined as inorganic non-metals, e.g., semiconductor materials and glasses as well as oxides, nitrides, carbides and borides.

Today, almost all information, including factual property data, is generated and can be collected – electronically. Key questions include:

- Do users want and need a single point of access for ceramics data?
- Are they willing to pay for it?
- What are the costs of maintaining such a resource?

Regarding database maintenance, keeping a data resource current is especially important, because materials change over time, especially in the case of ceramics, where a variation in starting materials and/or processing conditions can make a significant difference in the properties of the final product. Keeping updated with such changes is critical.

This paper includes a summary of discussions held at the 2010 meeting of the Interagency Coordinating Committee on Ceramic Research and Development on issues surrounding the development and maintenance of materials property databases, and access to the contents of these proposed repositories. These issues include the mutability of materials over time, the real and perceived value of materials property data, questions regarding the need and expense of evaluation and the proprietary nature of much of the property data and characterization of newly developed materials.

Participants in the ICCCRD meeting also discussed the recently added NSF requirement that its investigators submit a plan of how they will make their data available to the scientific community and how this requirement might impact materials property databases.

All of these issues are critical to understanding the future of materials databases and how access to them will evolve in the future.

Brief history of materials databases

An initial attempt to provide access to materials data took place in the 1960s with the development of the first crystallographic databases. In the late 1970s, the National Academy of Sciences conducted a study on the need for better access for data on aerospace materials.[2] The study recommended that the federal government investigate how to address this critical need.

In response, a 1982 international conference addressed the challenges associated with meeting the needs for computerized materials data systems.[3] Known as the Fairfield Glade conference, this seminal meeting strongly endorsed the development of a system that would include all of the following:

- Comprehensive materials properties data;
- A wide variety of data resources;
- Online access (predating the Internet and the world wide web);
- Multiple data providers;
- Subscription basis;
- Single point of access;
- Strong data integration standards;
- Cross-database searching; and
- Consolidated output.

In 1985, the Materials Property Council[4] began the National Materials Property Data Network that was based on a prototype system at Stanford University, funded by the National Institute of Standards and Technology (then the National Bureau of Standards), the Department of Energy and the Army.[5] Chemical Abstracts commercially operated the NMPDM until 1995. This network contained numerous databases on many different materials. Each new database involved a considerable investment by a different organization. Ultimately, this endeavor proved too costly to maintain and was abandoned.

One of the major factors that contributed to cost was that each type of materials property required a new database format. Adding a new database to the network was not an incremental cost. The experience of the NMPDM illustrates a major problem in the development and maintenance of a permanent single or small number of access points for ceramic materials. In contrast to many things, e.g., astronomical objects and crystals, which do not vary in composition over time, most materials of interest to engineers and scientists are constantly changing, with respect to overall composition, surface finish, properties and other factors. Collecting data on permanent entities is relatively straightforward. With time, a complete collection can be created.

However, in the case of materials, new ones are developed, while others are changed in composition and properties to improve performance or reduce cost. Therefore, to make a data resource relevant, it must be refreshed on a regular basis.

A particular problem with respect to ceramic property databases is a lack of specifications. The features of interest, e.g., microstructure, composition, purity and physical properties, may vary among different materials, even those with the same designation. A good illustration of this problem is aluminum oxide. The designation of a material as a 96 percent alumina means only that 96 percent of the bulk material is Al_2O_3. It says nothing about the other constituents, the processing procedures or the properties. Every "96 percent alumina" could be a different material with significantly different properties.

As new testing procedures are developed, test data from previous decades may not stand up to today's scrutiny. As an example, a comprehensive database on the fracture toughness and crack growth characteristics of inorganic glasses was established at NIST (then NBS) more than 20 years ago.[6] The database was created by collecting data from the literature. At the time, the collectors applied what was then considered to be quality evaluation. Today it is recognized that many of the test procedures used then have been improved upon. In other cases, it is known that the procedures could produce faulty results. Therefore, the test data are of little value. The situation with this database is not an isolated one.

The value of materials data

Some of the most significant uses of materials property data are in the design and manufacture of components. Materials data also are needed for failure analysis, processing improvement, product improvement and cost reduction. However, a significant consideration in investing in materials property data is its perceived value to the user. Routine uses of materials data may not be viewed as being worth the significant cost to develop, maintain and update a database.

Moreover, the value of data often is forgotten after its immediate use, especially if the data are used early in a planning and design process. As a consequence, there is an unwillingness to pay the real value of accessibility to materials data. On the positive side, some long-term investments have been made in collecting and providing access to materials data that continue to provide value. Prime examples of these investments are phase diagrams and crystallographic data.

Phase diagrams

NIST has partnered with The American Ceramic Society for more than 70 years in the collection, evaluation and dissemination of ceramic phase diagrams.[7] Systems include oxides, borides, nitrides, salts and electronic ceramics. There are 22,000 evaluated diagrams plus commentaries now available on CDs and in 21 bound volumes. An economic assessment study of the NIST–ACerS phase diagram program conducted in 1998 came to the conclusion that this program provided a benefit-to-cost ratio of approximately 10 to 1.

Crystallographic database

NIST partners with FIZ Karlsruhe to provide needed crystallographic data on inorganic crystals.[8] The database currently contains 132,000 peer-reviewed data entries, including atomic coordinates. NIST also provides a crystal database for inorganic and organic materials and a structural database for metals and intermetallic materials.

These databases contain fundamental materials property data, which remain of great value over decades. Performance property data, which is much more dependent on material composition and often varies over time as material composition changes, have not yet been successfully collected in long-term programs in the same manner.

Proprietary issues

Cost and performance competitiveness are always significant issues. In response, materials producers are continually redesigning, improving and changing commercial materials. In many instances, keeping proprietary data protected is critical for the economic success of the product or a company. Consequently, it can be concluded that significant quantities of data on the newest materials and latest materials advancements will be in large-scale databases only after a significant time lag.

Importance of quality

Ceramic property data can be obtained through Internet search systems and scanning of reports and published literature. An important question in this case is in the quality, reliability and provenance of these data. For example, which test procedures were used to obtain the data and were they conducted in a proscribed manner? In most instances, the individual needing the data is not an expert on the testing procedures and, therefore, cannot properly place a value on the quality of the data obtained.

Are data useful even if their pedigree is uncertain? Are any data, despite uncertainty in their quality, better than no data at all? That is, must data be evaluated by experts to be useful or valuable? Evaluation is expensive. It is not clear who would conduct such a task, nor is it clear who would be able and willing to pay for comprehensive evaluation services. Is self-evaluation a way to avoid the high costs of experts in a particular materials field? A publication by Munro[9] addresses such a question and provides a fundamental foundation for evaluation.

The International Centre for Diffraction Data[10] makes available a "quality mark" option, which allows the user to select only the most accurately determined patterns or to obtain all patterns regardless of quality. The quality marks reflect the type of X-ray diffractometer used, the degree of characterization of the chemical composition, an objective measurement of intensity and the spread in the peak positions.

Some trends for the future

With an increased use and development of nanomaterials has come an even greater need for sharing of critical materials property data. A new awareness of the importance of ready access to materials data and information has resulted in recent publications that discuss the current needs for materials informatics.[11] Most of the issues that apply to ceramics, including the funding and curating of databases, data mining techniques, linking databases to literature and providing single-point access, will be critical for nanomaterials as well. One of the crucial areas of need for information exchange is that related to environmental, health and safety issues posed by these new materials.

Regardless of the problems, information technology brings new capability to materials data users. The Social and Semantic Webs continue to evolve rapidly as mechanisms for materials data access and exchange. Modeling, simulation and knowledge discovery tools, as they become viable alternatives to materials testing, contribute new types of materials data, with different types of provenance and quality indicators. How that happens will be very interesting.

Data management plans for NSF proposals

For some time, NSF grantees have been expected to share primary data, samples, physical collections and other supporting materials created or gathered in the course of their work within a reasonable time at no more than incremental cost.[12] However, at the May 5, 2010, National Science Board meeting, it was announced that in the near future, scientists seeking NSF funding will be required to submit data management plans as part of their proposals.[13] The inclusion of a written plan is not completely new to NSF. For example, an older solicitation in the Biological Sciences Directorate[14] and many of the earlier versions of this solicitation, required a one-page plan. Effective January 18, 2011, inclusion of a plan for data management was not limited to a few solicitations – it became an NSF-wide requirement for all proposals.[15]

The motivation for this change is that science is becoming increasingly more data intensive and collaborative. Numerous disciplines often are needed to attack complex problems – sharing of data facilitates this process. Researchers communicate and collaborate through the processes of sharing data, software and publications. The ultimate goal is to have more effective research endeavors and make data available to future generations.

For the aforementioned BIO solicitation, data management plans required a description of the types of data to be produced, standards that would be applied, provisions for archiving and preservation, access policies and provisions, and a procedure or strategy for eventual transition or termination of the data collection after the NSF funding period. For 2011, the NSF-wide general requirements are specified in the Grant Proposal Guide. As well, several NSF units have provided more details about essential content and appropriate repositories (including databases)[16] to assist their communities and help define and inform about the review process.

In the coming years, NSF staff will have the opportunity to see the impact of the data management plans requirements. Some issues, such as open access (as applied to publishing) have not been addressed yet by NSF. However, as many researchers are aware, the National Institutes of Health has had a policy in place since 2008.[17] NSF is no stranger to data management – the Computer and Information Science and Engineering Directorate runs an Information and Data Management Program that "supports research and education activities fundamental to the design, implementation, development, management and use of databases, information retrieval and knowledge-based systems."[18] The expertise gained here can help guide changes. Within the NSF, a task force on data policies[19] continues to examine the many issues (e.g., definitions, technical considerations, repositories, international complexities, rights and legalities) and sometime in 2011 will produce a report of its findings and recommendations. Researchers should expect adjustments and refinements to NSF's data policy and its implementation.

The time has come

The complexity of materials and the existence of disparate materials classes have complicated the development of large-scale materials property databases. Although various sectors of the materials community have developed important data collections, the diversity and complexity of materials hinder strong business cases for a fully integrated materials data system, as envisioned nearly 40 years ago.

Despite the acknowledged need for data and protestations regarding the lack of access, the costs of developing and maintaining a comprehensive modern data system are higher than any one entity, i.e., company or government agency, can afford. A primary question is whether methods exist by which such a system(s) can be created based on shared costs. An additional question is what is the minimum level of expertise and sophistication needed to create worthwhile systems?

The past 30 years have seen tremendous changes in the technology available to provide new types of access to materials data. The same time period has provided many lessons about meeting the challenges of digitizing, managing and providing access to diverse types of materials data. Much of the progress made is the direct result of a robust planning process conducted in the 1980s that involved virtually every segment of the materials data community.

The time has come to plan for the future and assess where we are and where we need to go. This assessment requires understanding of what and how materials property data are cur-

rently available electronically on an international basis. The prioritized needs over the next decade for materials property data in government-funded research, as well as by commercial materials developers and manufacturers and academic researchers, must be determined. It is critical to determine viable business models that can be applied to establish a sustainable materials property data access approach. ■

Acknowledgments

LDM would like to thank H. Edward Seidel for sharing his thoughts and NSF's plans prior to the discussion of databases and data management at the ICCCRD meeting, and to Janice M. Hicks for her helpful comments on the manuscript.

Disclaimer

Any opinion, finding and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

About the authors

Stephen Freiman is president of Freiman Consulting. Lynnette D. Madsen is a Program Director in the Division of Materials Research at the National Science Foundation. John Rumble is Executive Vice President of Information International Associates Inc.

References

[1]"Integrated Computational Materials Engineering," *The National Academies Press*, Washington, D.C. (2008).

[2]"Mechanical Properties Data for Metals and Alloys–Status of Data Reporting, Collecting, Appraising and Disseminating," NDAB-NRC (1980).

[3]"Computerized Materials Data Systems," Office of Standard Reference Data, NIST.

[4]www.forengineers.org/mpc/index.html

[5]"Materials Information for Science &Technology (MIST): Project Overview," NIST Special Publication 726 (1986).

[6]www.ceramics.nist.gov/srd/summary/advmatdb.htm

[7]www.nist.gov/srd/nist31.htm

[8]www.nist.gov/data/nist3.htm

[9]"Data Evaluation Theory and Practice for Materials Properties," NIST Special Publication 960-11 (2003).

[10]www.icdd.com/

[11]"Materials Informatics," Edited by J.R. Rodgers and D. Cebon, *MRS Bulletin*, **31** [Dec] (2006).

[12]www.acpt.nsf.gov/bfa/dias/policy/dmp.jsp

[13]www.nsf.gov/news/news_summ.jsp?cntn_id=116928&org=NSF&from=news

[14]www.nsf.gov/publications/pub_summ.jsp?WT.z_pims_id=5337&ods_key=nsf09514

[15]www.nsf.gov/pubs/policydocs/pappguide/nsf11001/gpg_2.jsp#dmp

[16]www.acpt.nsf.gov/bfa/dias/policy/dmp.jsp

[17]http://publicaccess.nih.gov/policy.htm

[18]www.nsf.gov/funding/pgm_summ.jsp?pims_id=5210

[19]www.nsf.gov/nsb/committees/tskforce_dp_charge.jsp