Data in the 21st Century

For centuries, data have been distributed in print form. Over the past 50 years the traditional form of disseminating data in journal articles, handbooks and reference books has been expanded by the availability of electronic databases. Simultaneously, the increased use of modeling and simulation tools has led to an explosion in the need for data and the generation of new data. Retrieving data from traditional media or even electronic databases and preparing them for input into modeling and simulation tools can be time consuming and is prone to introducing errors. The white paper describing the Materials Genome Initiative (MGI) for Global Competitiveness [National Science and Technology Council, OSTP, Washington DC, June 2011; http://www.whitehouse.gov/sites/default/files/microsites/ostp/materials_genome_initiative-final.pdf] identifies data transparency, the availability of metadata, and innovative data transfer methods, together with the necessary modeling and simulation tools as key elements for achieving the goal of developing new materials at least twice as fast at a fraction of the cost.

Materials data range from fundamental data of the pure elements to processing dependent data of multi-component, multi-phase materials. A data infrastructure capable of serving multi-disciplinary communication between scientists and engineers needs to be developed. This data infrastructure will include data repositories that are capable of spanning the enormous space of materials data. Although the format for a repository must be formalized so that different individuals can access the data in the structure, it also must be flexible enough to accommodate extensions and new types of information while maintaining easy data retrieval. The format must be able to define a unique material in an unambiguous way such that the same material identifier can be used by other databases and tools searching for information on a specific material; this will permit links between these databases. In addition, material identifiers and phase names must also be unique and unambiguous. Examples of currently used identifiers for chemicals include the Chemical Abstracts Service (CAS) registry number [<http://www.cas.org/expertise/cascontent/registry/>] and International Chemical Identifier (InChI) [<http://www.iupac.org/home/publications/e-resources/inchi.html>]. An example for defining phase names is given by Schmid-Fetzer et al. [CALPHAD, **31**, 38-52 (2007)].

One possibility for a flexible data format is the use of extensible markup languages (XMLs). The use of XMLs would allow the almost limitless breadth of information, such as phase equilibria data, diffusivities, diffusion profiles, thermochemical data or mechanical property data, to name just a few, to be accommodated. Regardless of the choice of data format, the data stored should ideally include descriptions of the appropriate metadata. These metadata would not only include the material property being characterized, but would also include information describing how the individual datum was obtained. For phase data it is essential that the data are linked to the crystal structure of the phase.

Just capturing newly generated data, not to mention the task of capturing data from archival media, is a daunting task. This task cannot be carried out by one institution alone; communication and collaboration between a large number of institutions is imperative to achieve the goal of the MGI. The requirement that authors enter their data into a reference database as a prerequisite to the publication of an accepted paper has been very successful in a variety of scientific communities, such as genomics, crystallography, or thermodynamic data of organic systems. Before such an approach can be introduced for materials data, a system for depositing and storing these data must be developed. NIST (the National Institute of Standards and Technology) has been charged to target the development of reference databases and is currently working on defining standards for the collection data pertinent to modeling material properties and behavior, and on establishing protocols for the collection of data and their transfer to repositories. Once these tools are established, it is vital that materials science journals support the population of these repositories by requiring authors to enter their data as a prerequisite to the publication of an accepted paper.

Ursula R. Kattner
Associate Editor
Journal of Phase Equilibria and Diffusion