Big Data and Open Data are not closed topics

Data, in and of itself, is important, but it has little value unless it is needed and used for a purpose, such as to answer a specific question or provide input for decisions. Most available information is not without error, but in many cases, the goal of the information is not perfection but rather to allow the user the ability to launch into other areas of study or investigation. Thus, data should be captured in a way that is usable and accessible. If done well, Big Data and Open Data could lead to innovation, greater participation, collaborative sharing, and delivery of new products. But then there are the issues of maintaining databases, unequal sharing, privacy, and system compatibility.

These are some of the topics the Materials Research Society (MRS) and The Minerals, Metals and Materials Society (TMS) explored earlier this year when they established a committee representing various segments of the materials science and engineering communities and launched an "MRS-TMS Big Data Survey". The goal was to assess the current thinking on Big Data and Open Data within the greater materials community (see the Big Data article that appears in the September 2013 issue of *MRS Bulletin*).

"Big data is both a huge challenge and a huge enabler of discovery," said Eric Stach, Brookhaven National Laboratories and co-organizer of a Symposium X technical session on Big and Open Data for Materials Research at the 2013 MRS Fall Meeting (http://www.mrs.org/fall-2013-big-open-data/). "If you are able to capture, analyze, and understand the information, there are tremendous opportunities in dramatically accelerating the rate of discovery in material research."

The MRS/TMS survey came on the heels of the Open Data Policy that was issued by the US White House this year requiring federal agencies to

- use open licenses—make data public in such a way as there are no restrictions on copying, publishing, distributing, transmitting, adapting, or otherwise using the information for non-commercial or commercial purposes;
- use standard metadata—"data about data"—to tell users where each data set comes from, when it was collected, and what its quality is;
- support interoperability (making it possible to analyze one data set with another) and information accessibility (making data usable in the first place);
- build an inventory of all of the agency's data sets and publish a list of all the ones that are open to the public; and
- protect privacy and confidentiality, and keep data secure (www.opendatanow.com).

The US Office of Science and Technology Policy (OSTP), part of the Executive Office of the President, issued a policy memo directing all federal research agencies to develop and implement open access plans over the next two to three years (http://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf).

"The biggest problem I see with a big data mandate is the ability to disseminate the variety of data in a useful format to all interested parties," said one respondent to the MRS-TMS survey. Another respondent felt that open access and big data will undermine discovery, sound publications, entrepreneurship, and job creation. "It is somehow the promise that a big pile of data and a lot of computing power would be sufficient to gain new insights. This promise will not be fulfilled," said another.

The MRS-TMS survey aimed to identify areas of possible agreement and contention regarding big data and open data. One of the questions in the survey asked whether the researcher would participate if data sharing was "encouraged" as a term and condition of funding or publishing, assuming the proper safeguards were in place. Seventy-four percent of the respondents to the survey said they would participate, with conditions regarding confidentiality, funding, organizational policies, and the type of research.

"I don't think this will have a big negative impact on government funded collaborative projects between companies and universities, as in those cases, companies usually have accepted that data being developed is going to end up public and shared eventually," said Dane Morgan, a member of the MRS Survey Committee and co-director of the Wisconsin Materials Innovation Institute at the University of Wisconsin–Madison. The institute was selected last summer as a partner in the federal government's Materials Genome Initiative for Global Competitiveness. This government initiative aims to double the speed in which new materials are discovered, developed, and manufactured.

"While an initial response might be that industrial participation in federally funded programs would decrease, if appropriate and agreed-upon practices and safeguards to maximize mutual benefits for universities, government labs, and industrial partners were put in place, industrial participation could increase significantly," said Laura Bartolo, a member of the MRS Survey Committee

and professor and director of the Center for Materials Informatics at Kent State University in Ohio.

"With industry, they are going to be quite reluctant to share information that provides them with a competitive advantage in their business. And that's not going to change," said David P. Norton, vice president for research in the Office of Research at the University of Florida. Norton is also a co-organizer of the Symposium X technical session on open data at the 2013 MRS Fall Meeting that will include a panel discussion. "The advantage of the ability to share data/open access benefits more of those projects that don't have commercialization opportunities. That gets complicated with IP and 'trade secrets'—things that would provide them with competitive advantages."

Regarding open access, the top three motivations that survey respondents cited as encouragements for sharing their data on an open-access basis were (1) increased visibility of research/work, with 72% of respondents listing this as a motivation; (2) the opportunity to receive feedback from others about the data (67%); and (3) the opportunity for others to analyze the data (and potentially make other discoveries as a result) (54%).

"More access for more people leads to faster materials research and development," said Morgan. This could also give "opportunities for mining correlations we have not seen before … and a reduced reliance on other researchers' interpretations as more of their data is accessible."

Norton agreed, "There is an opportunity for better science by engaging a broader community. Data from a number of sites allows you to do remarkable science that you could not do isolated."

The top impediments identified by survey respondents were (1) the proprietary/restricted nature of their data, with 59% viewing this as a barrier; (2) the intellectual property rules within their organization/business (54%); and (3) the fact that their data were stored within a propriety data format (42%).

"You cannot make data available/searchable without proper infrastructure and management," said Norton. "There's also the impact on intellectual property. It's not an unmanageable thing, but the correct guidelines have to be there for the institutions to protect their property."

Bartolo said, "One area [deterring the open exchange of data] is the variability among researchers for describing and organizing their data, which can impede correct interpretation, comparison, exchange, and reuse of data and which community-driven standards or best practices could help mitigate. Another area includes intellectual property, export control, and ITAR [International Traffic in Arms Regulations] restrictions." Bartolo also suggested that workshops and reports can bring together stakeholders to review current practices and help develop proposed recommendations. "Big data brings along another set of technical issues, such as inadequate technical capabilities and capacity of computer systems, data storage, and communications networks."

"We will spend more time in preparing and sharing data for public dissemination, possibly in new forms and to a much larger extent," said Morgan. He expressed concerns about errors associated with how people use raw or low quality data. "There may be some learning curve for the community of materials researchers in this area of using more raw data." Morgan also points out that there could be counterproductive pressures associated with forcing more sharing, noting that "today people commit time and resources to generating large data sets to obtain an advantage over competition, and they may feel it is no longer worth the effort if they cannot keep it private."

When asked what types of data should be made available for open access, the responses varied. Some felt only peer-reviewed data should be presented. Other participants thought experimental data could be included if information is given on how the information was derived. Another respondent felt "it may be best suited (at least initially) for older data sets that have been vetted by experts and well-referenced in the literature. Obviously, researchers would prefer the latest, cutting-edge information. Unfortunately, this type of data is likely the most sensitive and subject to restrictions."

Norton and Morgan agreed with the survey responses favoring open access to federally funded research published in peer-reviewed journals. Morgan takes it a point further: "I would also include more raw data and detailed data that we would typically never publish (e.g., all 10,000 steps of a molecular dynamics simulation)." He draws the line on work that is classified, for example, due to strategic/military importance.

The difficulty in open access, said Norton, "are data that have not been distilled. Who's paid to do the data? Does it affect future publications for students working on their dissertations? For state-funded institutions, we have a motivation to encourage economic development. It's a balance."

The path forward with big data and open data for materials advances requires careful planning and well-designed road maps to offer the highest likelihood of success. However, the course is neither simple nor straightforward. Revolutionary possibilities lie ahead, and the last word on the access of information has not yet been spoken.

Lori A. Wilson