Survey on Big Data gathers input from materials community

"Big Data" and "Open Data" are terms increasingly heard when referring to the vast amount of digital data that are available today and are topics of great interest in the scientific community. The outcome of discussions on these topics will play a major role in the progress of scientific enquiry in the future. Any conclusions could also affect the policies adopted by government and private funding agencies, publishers and editors, universities and private research institutions, and individual research groups worldwide. To set the stage for these discussions, the Materials Research Society (MRS) and The Minerals, Metals and Materials Society (TMS) find it important to identify key areas of possible agreement and disagreement at the outset.

So earlier this year, MRS and TMS established a committee to develop a survey to assess the current thinking on these key topics within the greater materials community. The diverse committee included members representing various segments of the materials science and engineering community, and in the spring of 2013, they launched a 25-question "MRS-TMS Big Data Survey."

"The MRS-TMS Big Data/Open Data Survey provides a needed baseline foundation for building community discussions and future surveys on data and access to accelerate the discovery of new materials," said Laura Bartolo, survey committee member and professor and director of the Center for Materials Informatics, Kent State University.

In mid-February, the White House Office of Science and Technology Policy (OSTP) issued a memorandum that stated that in order to achieve the Administration's commitment to increase access to federally funded, published research and digital scientific data, federal agencies investing in research and development must have clear and coordinated policies for increasing access to digital data sets. This "allows companies to focus resources and efforts on understanding and exploiting discoveries." According to the memorandum, "this will allow wider availability of peer-reviewed publications and scientific data in digital formats and will create innovative economic markets for services related to curation, preservation, analysis, and visualization" (http://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf).

Nicola Marzari, survey committee member and professor at École Polytechnique Fédérale de Lausanne, Switzerland, said, "Open-source data is key to the accountability and reproducibility of scientific research, and a duty for any publicly funded project. In order to make this possible, we need to work on data standards and on data repositories. Data infrastructure, as well as software infrastructure, is these days as important as brick-and-mortar facilities."

Demographics

Of the 25 questions in the survey, 23 included standard multiple choice answers that participants could choose, as well as an "other" option so they could explain answers that did not fit into one of the standard categories. Comment boxes were appended to several survey questions. The last two questions on specific developments were open-ended and gave the respondents an opportunity to provide feedback on the evolution of tools and data, as well as their thoughts on the broad areas of big data and open data.

There were 675 respondents when the survey was closed on June 3, of which 73% completed the survey. The other 27% responded to most of the questions but did not finish the survey. Responses from all participants are included here, whether or not they finished the survey. Sixty percent of respondents were from the United States, followed by Germany, the United Kingdom, India, Japan, and Canada, with about 3% each. Overall, 64% were from North America, 17% from Europe, and 13% from Asia.

Of the respondents, 47% were from academia, 27% from industry, and 14% from government. The remainder were from small business, nonprofit professional organizations, and other unspecified groups. Forty percent identified themselves as scientists/engineers, 26% as university professors/faculty, 11% as executive/management, 7% as postdoctoral researchers, and 8% as undergraduate and graduate students.

Survey Question: What scientific/technical databases and data mining tools would be most useful to you if they could be created?

Physical and thermal properties	87.0%
Chemical properties	61.5%
Microstructure, photo records	57.8%
Static mechanical properties	52.4%
Engineering/processing details	50.6%
Electronic properties	42.8%
Cyclic or dynamic mechanical properties	41.9%
Environmental and corrosion	37.2%
Other	6.5%

https://doi.org/10.1557/mrs.2013.225 Published online by Cambridge University Press

The remainder were from the press, sales/marketing, retired, unemployed, and other unspecified groups. Forty-eight percent of respondents identified themselves as experimentalists, 32% applied/development, 13% computational, and 4% theorists.

Survey results

Regarding the use of software tools and databases, 82% of respondents use software tools for data analysis and 64% for data processing. Approximately 50% use such tools to visualize data and for simulations. Forty-three percent of respondents use materials databases for experimental design, and 53% for interpreting/modeling experimental data. Also, 53% use materials databases as input for calculations, 31% for model validation, and 25% for data mining. Respondents use databases for materials selection (34%) and for product engineering design (20%).

For the types of desired data, 87% of respondents said a physical and thermal property database would be useful, while 62% want a chemical properties database. The results show a need for data over a broad range of areas, including mechanical properties, electronic properties, microstructures/photographs, and engineering details (see Table).

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 (72%), (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%). Conversely, the top impediments identified by survey respondents were (1) the proprietary/restricted nature of their data (59%), (2) the intellectual property rules within their organization/business (54%), and (3) the fact that their data was stored in a proprietary data format (42%).

The majority (74%) of respondents said that if data sharing were to be "encouraged" as a term/condition for funding or publication, they would participate. To ensure the quality of shared data, approximately 46% said that metadata standards and protocols should be established standards, while 30% preferred community-driven standards.

Regarding responsibility for enforcement of data management and sharing policies, including date of data release, at the funding stage, 69% said the funding agency should enforce compliance of data management and sharing policies. At the publication stage, 21% said the publisher should enforce compliance of data management and sharing policies, while 14% thought it should be the funding agency. Other responses included journals, institutions, and the individual researcher/author. Responses were closely divided on whether data, including data not used for the publication, should be released at the date of publication, with 47% saying yes and 43% saying no. "Other" responses suggested that only supporting data should be required, or that it is situation-dependent.

Fifty-seven percent of respondents stated that additional actions are needed to ensure that publications sufficiently describe the experimental and computational details required to reproduce data. Suggestions for these actions include establishing standards and guidelines, expanding experimental/modeling sections, verifying data through peer review, and providing details/code required for reproducibility.

Open-ended comments

In the open-ended comments, when asked to identify any areas or specific developments seen as critical to the evolution of data, tools, and cyberinfrastructure for materials science in the near or long terms, the majority of concerns could be grouped into five categories:

- *The need for better data, not more data.* One respondent said having large volumes of data is meaningless if they cannot be analyzed and useful information extracted. By sharing all data, there is the fear of including poor or bad data, resulting in researchers spending a lot of time looking at wrong data. "The problem with today's literature is not the shortage of open data, but the dumping of inadequately curated and analyzed data into literature," said one respondent. Thus, standards must be in place for formatting, data sharing, and gatekeeping to ensure quality. One suggestion in dealing with this was ongoing and careful validation of computational data with experimental and real-life data sets. This leads to the next concern.
- *Finding staff time and funding to support these policies.* Full implementation and compliance with OSTP policies may require more

https://doi.org/10.1557/mrs.2013.225 Published online by Cambridge University Press

resources than are available in a time of restricted budgets. One respondent felt that the construction of new extended, open databases was not in the interest of good, high-quality science. Time spent publishing data is time that should be spent doing science, one respondent said. The cost of making data publicly accessible is still unknown. Thus, there is concern for financial compensation for industrial investment in data generation, software tools, storage, and support of access.

- *Compatibility and accessibility of information.* Many individual data management systems have limited functionality and are not easily linked to other systems, making collaboration difficult. Databases will need to be searchable, extensible, and relational and have tools that work across all disciplines. Some requirements may be specialized search engines in conjunction with standardized metadata, standard format for submission, software to handle the submissions, and standardized reporting formats and terminology. The ability to disseminate the variety of data in a useful format to all interested parties is important, and "compatibility will always be a challenge when access is expected on a global scale," said one respondent.
- *Open international sharing of data.* A concern expressed is that if researchers from a select few countries openly share information, these researchers are at a disadvantage because they are giving information away and possibly not receiving it in return. There will be a need to collaborate with other materials organizations around the world to have a common sharing data platform, a respondent asserted. The United Kingdom, for example, recently implemented open access policies for peer-reviewed publications, and the European Commission has plans to expand its open access requirements.
- *The understanding of Intellectual Property issues.* Some concerns regarding sharing data included plagiarism, data duplication, and violations of patents and copyrights. It was suggested that "ownership" should be defined and security questions addressed.

Wrap-up

"The survey shows that we are a community that widely uses available software and data, wants more access, and is ready to participate in sharing tools and data at a much greater level. This suggests that as structures are evolved to enable more sharing, we will see an explosion of accessible tools and data and, following from that, increasingly efficient materials discovery, optimization, and utilization," said Dane Morgan, survey committee member and associate professor of materials science and engineering, University of Wisconsin.

This survey clearly showed that the members of the materials community have differing opinions on big data and open data, and a number of discussions need to occur before specific policies can be recommended or implemented. Ultimately, the goal is better processes and tools for sharing information to advance the overall field of materials. This survey was the first step in this endeavor.

Lori A. Wilson

https://doi.org/10.1557/mrs.2013.225 Published online by Cambridge University Press