Construction of Integrated Materials Data System for Data-driven Materials Research

Yibin Xu

National Institute for Materials Research, Tsukuba Ibaraki 3050047, Japan
Xu.yibin@nims.go.jp

Big data based artificial intelligence has been greatly expected to change the style of materials research, and improve the efficiency and decrease the cost of materials development. However, in spite of the great efforts done on data collection and database construction since 1880's, data shortage is still the bottleneck of today's data-driven materials research.

Data integration has been addressed as one of the key issues of materials data system for decades. In recent years, some data formats and protocols have been proposed for data exchange between different data resources. Nevertheless, material identification is still a confusing problem, since lack of common descriptors for all materials fields. In this presented work, based on the statistics of NIMS Inorganic Materials Database AtomWork-Adv, we proposed a set of descriptors to define a substance, and created a substance dictionary containing more than 158,000 substances. With this dictionary, users can identify their own materials at substance level, and make links to the crystal structure and property data in AtomWork-Adv. We also show an example of data structure for multiphase and composite materials and how to link them with the substance data.

Addressing the data shortage problem, we analyzed the distribution of available property data versus substance. This helps us to understand the current situation of data availability and set up plans of data generation. We also show that a key to small data issue is to leverage the correlations between properties. Since properties dominated by same physical and chemical factors tend to have strong correlationship, it gives us a chance to use the property with sufficient data as a substitute of that with a few data available. Several examples of our data-driven researches with small data set are introduced.