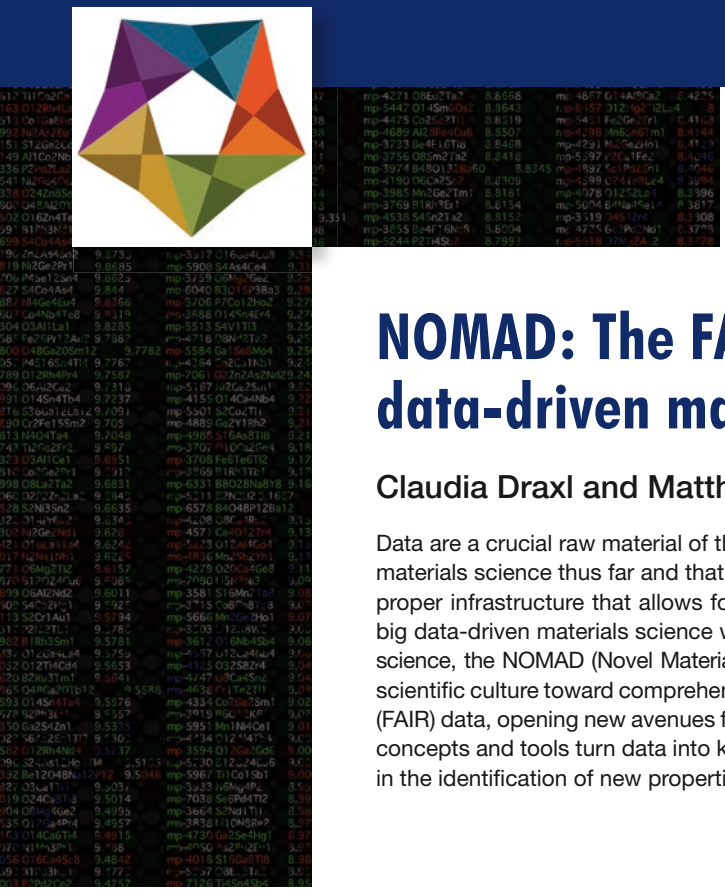


NOMAD: The FAIR concept for big data-driven materials science

Claudia Draxl and Matthias Scheffler

Data are a crucial raw material of this century. The amount of data that have been created in materials science thus far and that continues to be created every day is immense. Without a proper infrastructure that allows for collecting and sharing data, the envisioned success of big data-driven materials science will be hampered. For the field of computational materials science, the NOMAD (Novel Materials Discovery) Center of Excellence (CoE) has changed the scientific culture toward comprehensive and findable, accessible, interoperable, and reusable (FAIR) data, opening new avenues for mining materials science big data. Novel data-analytics concepts and tools turn data into knowledge and help in the prediction of new materials and in the identification of new properties of already known materials.

Introduction

The discovery of improved and novel—not just new—materials or unknown properties of known materials to meet specific scientific or industrial requirements is one of the most exciting and economically important applications of high-performance computing (HPC) to date. The convergence of theoretical physics and chemistry, materials science and engineering, and computer science into computational materials science enables the modeling of materials, both existing materials and those that can be created in the future, at the electronic and atomic levels. This also allows for the accurate prediction of how these materials will behave at the microscopic and macroscopic levels, and an understanding of their suitability for specific research and commercial applications. Computational high-throughput screening initiatives,*[1–5] significantly boosted by the US Materials Genome Initiative,[6,7] meet this issue by computing the properties of many thousands of possible materials.

When looking more closely, however, one realizes that such studies have been inefficiently and ineffectively exploited so far, as only a tiny amount of the information that is contained in all the computed data is being used. This applies to high-throughput screening as well as to individual theoretical and experimental investigations. Unfortunately, besides a few numbers, tables, or graphs that appear in resulting publications, the wealth of other information contained in the full research work is typically disregarded or even deleted.

Changing this situation and fully realizing comprehensive data sharing is slow, and characterized more by lip services from science policy and funding agencies than by real commitments and support.[8] This has slowed possible progress of scientific advancements. In this context and in the area of computational materials science, the NOMAD (Novel Materials Discovery) a European Center of Excellence (CoE),[9] whose computer and storage are at the Max Planck Computing and Data Facility in Garching/Germany, has assumed a pioneering role, considering all aspects of what is now called findable, accessible, interoperable, and reusable (FAIR)†,[10] handling of data: Data are findable for anyone interested; they are stored in a way that makes them easily accessible;[11] their representation follows accepted standards,[12,13] and all specifications are open—hence data are interoperable.‡ All of this enables the data to be used for research questions that could be different from their original purpose; hence data are repurposable.§

We illustrate the latter with an example. Let's assume a research team has investigated TiO_2 for heterogeneous catalysis where TiO_2 is an important support material. The results published in a research journal are typically not useful for researchers who are interested in the same material,

Claudia Draxl, Humboldt-Universität zu Berlin, and Fritz-Haber-Institut Berlin, Germany; claudia.draxl@physik.hu-berlin.de
Matthias Scheffler, Fritz-Haber-Institut Berlin, and Humboldt-Universität zu Berlin, Germany; scheffler@fhi-berlin.mpg.de
doi:10.1557/mrs.2018.208

© 2018 Materials Research Society
https://doi.org/10.1557/mrs.2018.208 Published online by Cambridge University Press

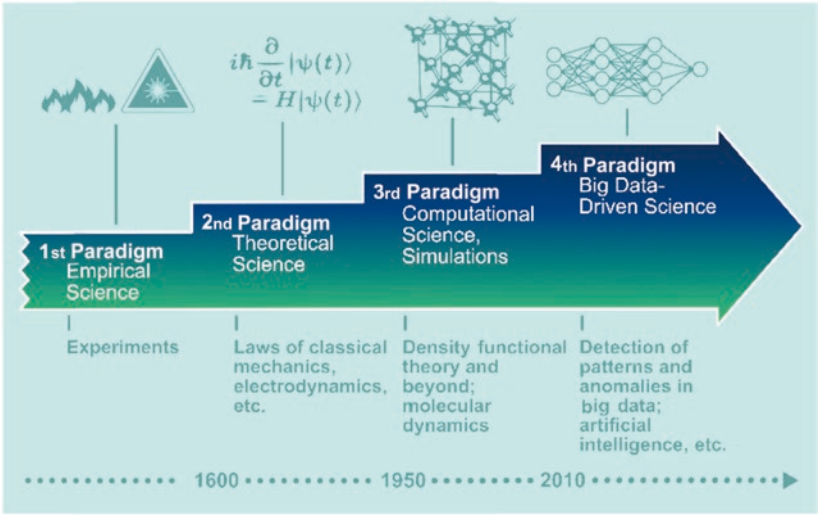


Figure 1. Development of the four materials science and engineering paradigms.

but in a different context. Having all data in a normalized and code-independent form, as they are available in the NOMAD Archive,[14] allows one to use it in a versatile form. This data can be as useful for photovoltaics in inorganic/organic hybrid materials as well as for research on pigmentation. Recall that TiO_2 is behind the white color in paints and toothpaste.

Today, big data and artificial intelligence are revolutionizing many areas of our lives and the sciences. Materials science is not an exception (see References 15–28). In fact, data-driven science is becoming the fourth paradigm of materials research, as sketched in **Figure 1**.** While experimental investigations have been carried out since the Stone and Copper ages, scientists of the late 16th and the 17th century started to describe physical relations by equations. Thus analytical equations became a central instrument of theoretical physics (symbolized by the Schrödinger equation in Figure 1). Obviously, with the advent of the second paradigm, the first one, empirical and experimental science, did not become obsolete, but rather was complemented by an important, novel methodology. The 1950s and 1960s marked the beginning of computational materials science and simulations, the third paradigm. Within this framework, computer-based experiments and simulations became possible, with the corresponding results being analyzed und interpreted like the experimentally measured ones.

The vision of NOMAD was and is to establish the fourth paradigm[29] in computational materials science (Figure 1) with the recognition that big data contain correlations, reflected in terms of structure and patterns in the data that are not visible in small data sets. Furthermore, we acknowledge that many properties of materials cannot be described by a closed mathematical expression as they are determined by several multilevel, intricate theoretical concepts. The various methods of artificial intelligence (e.g., machine learning,[27] compressed sensing,**,[26] subgroup discovery[28]) can identify such structures and transform them into useful information (**Figure 2**).

Thus, a key aspect of NOMAD's vision is to draw "materials maps" that allow us to find materials that are useful for a particular application in the "structural and chemical compound space," as sketched in Figure 2. Works like those of Phillips and Van Vechten[30,31] and Ashby plots[32] are role models, although these "materials maps" employ "(x, y) coordinates" that necessitate some prior knowledge about the described material. For atoms, the Periodic Table of the Elements is a fascinating early role model for our vison for such maps, where totally unknown elements were identified in terms of white spots in the table, and even their rough chemical properties could be predicted. Amazingly, this table was developed before any understanding of quantum mechanics and the shell structure of electrons. Materials maps are surely much more complex and multidimensional. The crucial challenge in this context is finding the actuators behind the material properties and functions (i.e., the proper descriptors). These need to be based on atomic properties and "collective" quantities that are easier to determine than those to be predicted. Turning this vison into reality requires novel approaches, and we need extensive data management. Similar to the establishment of a stable Internet, we now need to install a comprehensive big data infrastructure.

Finally, we stress that the success of novel data-mining tools strongly depends on the data quality.[33] The NOMAD Laboratory[9] has set out to address all these aspects of computational and data-driven materials science, providing a true platform for open materials science. It starts with the NOMAD Repository,[34] by now the world's largest raw-data

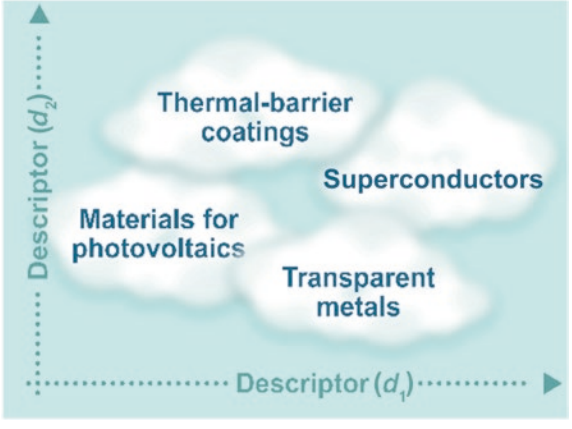


Figure 2. Big data contain correlations and structures that are invisible in small data sets. Finding descriptors that determine a specific property or function of a material is a crucial challenge. Once they are in place, machine learning can be used on the data to eventually draw "maps of materials."

https://doi.org/10.1557/mrs.2018.208 Published online by Cambridge University Press

collection of its kind. In the NOMAD Archive,[14] the raw data are brought to a normalized form. The NOMAD Encyclopedia[35] is the "face" of the project, displaying the content of the vast amount of data, thereby also using advanced visualization tools. Finally, the NOMAD Analytics Toolkit demonstrates novel data-mining methodology for data-driven research. The overall concept of the NOMAD CoE is sketched in **Figure 3** and is described in more detail next.

The NOMAD Laboratory

The NOMAD Laboratory[9] is a European CoE that was established in 2015. Eight complementary research groups along with four high-performance computing (HPC) centers, in different places in Europe form the synergetic core of this CoE. This group also reflects that the CoE is part of the Psi-k,[‡‡,36] CECAM,[§§,37] and ETSF***,[38] communities.

In short, the NOMAD CoE creates, collects, processes, stores, cleanses, and visualizes computational materials science data computed by the most important materials science codes[††] available today. More importantly, the NOMAD CoE develops innovative tools for mining this data in order to find structure, correlations, and novel information that could not be discovered from studying smaller data sets. The big picture is to advance materials science by enabling researchers in basic science and engineering to understand materials data, identify new materials and physical phenomena, and thus advance basic science and help industry to improve existing, and develop novel, products and technologies.

The NOMAD Repository

The first level of the NOMAD CoE is the NOMAD Repository. It currently contains the input and output files from several million high-quality calculations; the data volume made available through the NOMAD Repository is rapidly increasing. In fact, the computational materials science community uses millions of CPU hours every day in HPC centers worldwide. The NOMAD data collection comprises calculations that have been produced with any of the leading electronic-structure software codes and increasingly also with codes from quantum chemistry. Presently, NOMAD supports about 40 codes (see also the NOMAD archive section next), and less-frequently used codes will be added on demand. Thus, the NOMAD concept and praxis are orthogonal to other data collections, because the NOMAD Repository is not restricted to one or two selected computer codes or closed research teams but serves the entire materials research community with its ecosystem of very different computer codes.

The NOMAD Repository contains not only plenty of calculations performed by individual researchers from all over the world, but also data from the most important computational materials databases worldwide. Examples include AFLOW,[39] directed by S. Curtarolo at Duke University, OQMD,[40] directed by Ch. Wolverton at Northwestern University, and the Materials Project,[41] directed by K. Persson at the University of California, Berkeley. As such, NOMAD copes with the

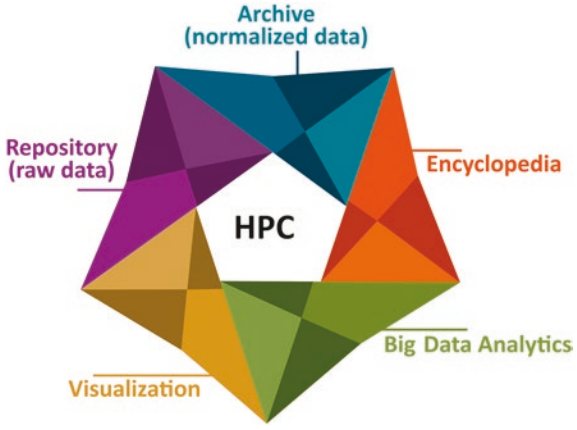


Figure 3. Structure of the Novel Materials Discovery Center of Excellence (NOMAD CoE).[9] More information can be found at https://youtu.be/yawM2ThVlGw (the NOMAD Laboratory CoE, 5-30-2017). Note: HPC, high-performance computing.

increasing demand for storing scientific data and making them available for longer periods, as required by many funding agencies worldwide. NOMAD keeps scientific data for at least 10 years, for free. In this way, NOMAD helps research groups to organize their data for their own groups and to share and exchange their results between two or more groups, making it easier to recall what was actually done some years ago. The NOMAD Repository is the only repository in materials science so far to be recommended by Nature journal *Scientific Data*.[42]

Results are accepted in their raw format as produced by the underlying code. The only conditions are that the list of authors be provided, the code and version be retrievable from the uploaded files, and the input and output files be complete. Data can be restricted to the owner or made available to other people (selected by the owner). After a maximum period of three years, all data become open access. For downloading data, registration is not required. As previously noted, the NOMAD concepts are 100% FAIR.[†,‡,10] The NOMAD Repository has been online and open since the beginning of 2014 and has enabled a cultural shift toward open data and thus open science[43] in the computational materials research community. **Figure 4** shows the numbers of total-energy calculations produced by different codes found in the NOMAD Repository and Archive (as of March 15, 2018). For a quick introduction to the NOMAD Repository, we recommend watching the short movie listed in Reference 44.

The NOMAD Archive

As the NOMAD Repository data is generated by many different computer codes, it is very heterogeneous. We have developed ways to convert the existing open-access data of the NOMAD Repository into a common, code-independent format, by developing numerous parsers and creating the NOMAD Archive. The NOMAD Archive consists of the open-access Repository data, converted to a normalized form. This ensures that data

https://doi.org/10.1557/mrs.2018.208 Published online by Cambridge University Press

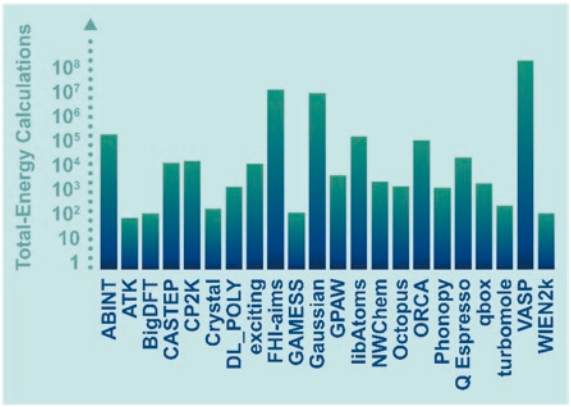


Figure 4. The Novel Materials Discovery (NOMAD) Laboratory supports all important codes in computational materials science. The figure shows the number of uploaded open-access total-energy calculations at the NOMAD Repository* as of March 15, 2018. The abscissa shows the various codes with more than 80 uploads. The total number of open-access total-energy calculations at the NOMAD Repository is more than 50 million, corresponding to billions of CPU-core hours.

from different sources can be compared and, hence, collectively operated upon by various NOMAD tools. A clear and usable metadata definition is a prerequisite for this normalization step to a code-independent format.

The development of the NOMAD Meta Info[14] was indeed a challenge. When NOMAD started, practically no metadata existed for describing materials data, but they are of crucial importance since physics, chemistry, and materials science may use one term for different quantities or different terms for one and the same property.

The NOMAD Encyclopedia

Knowledge and understanding of materials is based on their characterization by a variety of measured and computed properties. This includes structural features, mechanical and thermal behavior, electronic and magnetic properties, response to light, and more. On the computational side, the NOMAD CoE has created a data infrastructure for not only collecting and sharing data, but also indicating what this data contains. Indeed, the variety of data uploaded to the NOMAD Repository contains a lot of information that has not even been fully explored by the researchers creating the data. The NOMAD Encyclopedia is a web-based infrastructure that makes many millions of calculations accessible. Its graphical user interface provides a materials-oriented view of the computational materials data of the NOMAD Archive, displaying all properties of a given material which have been computed all over the world. In other words, it represents a user-friendly, public access point to the extensive knowledge contained in the NOMAD Archive. It not only helps investigators search for the properties of a large variety of materials, it also allows one to directly observe the spread of results, for instance, the impact of a density functional on a given feature. Thus far, structural and electronic properties and thermal behaviors have been processed, both for bulk materials and low-dimensional systems. The Encyclopedia will soon be able to handle molecules, surfaces and adsorbate systems, response to external excitations, elastic properties, Fermi surfaces, molecular dynamics, and more.

The Encyclopedia also provides a material classification system, links to external sources, and has an error-reporting tool for problematic data. Should there be a data set or a graph that does not appear to be correct, a user can contact us with the help of a simple menu.

The NOMAD visualization tools

Seeing helps understanding. NOMAD has developed an infrastructure for remote visualization of multidimensional NOMAD data. We provide a centralized service that enables users to interactively perform comprehensive data visualization tasks on their computers without the need for specialized hardware or software installations. It allows researchers to conveniently perform graphical analyses of complex and multidimensional time-dependent data from electronic-structure simulations, together with molecular structures. There is special focus on virtual reality (VR) for interactive data exploration. Users have access to data and tools using standard devices (laptops, smartphones). Such VR enhances training and dissemination, and is especially successful when presented to the general public. As an example, 360° movies can be watched with simple Google cardboard glasses, as demonstrated for CO_2 adsorption on CaO[45] and excitons in LiF.[46] The latter, being six-dimensional objects, cannot be easily visualized in an insightful manner otherwise. Taking the position of an electron or a hole, VR allows viewers to inspect the space of its counterpart.

The NOMAD Analytics Toolkit

With the data of the NOMAD Repository and Archive at hand, we now address the question of how this data can be turned into knowledge and understanding. We should emphasize that the amount of available data is huge (billions of results), but compared to the immensity of the number of possible materials and the many intricate processes that determine materials properties, coverage of the structural and chemical compound space is still shallow.

Interestingly, this immensity in number of possible materials becomes much smaller when the focus is on selected properties or functions (e.g., systems with a certain bandgap and effective mass that are stable under ambient conditions). Thus, we are looking for some needles in a haystack, and if our tools are too approximate, we may describe the hay but we may miss the needles. Our aim is to develop big data analytics tools that will help to sort all of the available materials data to identify trends and anomalies and to build maps that also cover materials that have not yet been synthesized, as sketched in Figure 2.

Machine-learning approaches are nonlinear fits of a large pool of data. They will work only when there is sufficient data.

https://doi.org/10.1557/mrs.2018.208 Published online by Cambridge University Press

However, for most materials properties, this is often not the case, and the question has been raised whether there will ever be enough data. The approach will work on less data if based on a clever descriptor (a set of descriptive parameters) that ensures that the data is arranged in a somewhat smooth manner.[24–26] Finding the correct descriptor will require domain knowledge, particularly when a highly accurate and predictive description is required. Progress in this direction has been significant in exploiting the mentioned sparsity by compressed sensing,[24–26] but in general, the systematic search for descriptors for materials science is still in its infancy.

In this spirit, the NOMAD CoE is developing an "Analytics Toolkit." The overarching topics that are currently addressed are (1) crystal structure prediction, (2) scanning for good thermoelectric materials, (3) finding better materials for heterogeneous catalysis, (4) searching for better materials for optoelectronics and photovoltaics, (5) analyzing alloys and their plasticity, and (6) predicting topological insulators. A number of prototype applications of our data-mining approaches are already publicly available.[47]

We can illustrate the concept of finding descriptors and building "maps of materials" by a recent example that employed compressed sensing for the descriptor identification. Compressed sensing originates from signal processing and finds a low-dimensional representation for a complex signal. In materials science, it has been used to identify the key physical "actuators" that are behind materials properties, and it identifies a few leading descriptor equations out of a huge number of candidates.[26] In recent work, Acosta and co-workers[48] used the above approach to build a materials map for two-dimensional (2D) honeycomb structures in order to analyze and identify 2D topological insulators (also called quantum spin Hall insulators [QSHIs]). The authors calculated 220 functionalized honeycomb lattices that are isoelectronic to functionalized graphene. These materials are built of Group IV or III–V or IV–VI elements, and all atoms are bonded to a Group VII element. Besides confirming the QSHI character of known materials, the study revealed several other as yet unreported QSHIs. Using a recently introduced method called sure independence screening and sparsifying operator,[26] the authors then offered 10 million candidate descriptors to identify the best low-dimensional descriptor. **Figure 5** shows the corresponding "materials map" defined by the 2D descriptor. Analysis of these descriptors (not shown here) yields fundamental insights into the mechanisms driving topological transitions. Furthermore, the map predicts several new QSHIs that were not part of the calculated materials.[48]

Summary and outlook

The growth of data from simulations (and experiments) is expanding beyond a level addressable by established scientific methods. The so-called "4 V challenge" of big data–volume (the amount of data), variety (the heterogeneity of form and meaning of data), velocity (the rate at which data may change or new data arrive), and veracity (uncertainty of the data quality)—is clearly becoming eminent in materials science. Controlling these massive amounts of data sets the stage for explorations and discoveries. Novel data-mining technologies can find patterns and correlations in data that cannot be seen in small data sets or standard tools. As such, data-driven materials research is adding a new research paradigm to our scientific landscape. All this is at the heart of the NOMAD CoE.

Handling experimental data in a similar fashion as computational data—the logical next step—will make the data challenge even more difficult. Today, experiments produce terabytes of data per day even in mid-sized facilities. Taking transmission-electron microscopes as an example, not only can detectors acquire thousands of data sets per second with millions of data points at once, they can also operate in parallel for multisignal acquisition. This leads to data creation at typical rates of 30 GB per second—orders of magnitude faster than existing data processing and analysis tools can cope with. If this mass of data is to be effectively captured, managed, navigated, and exploited, novel HPC solutions for high-performance data processing, data storage, data analytics, information retrieval, and user experience are essential. Notably, experimental data stem from various measurement techniques and instruments with different resolutions and other experimental parameters. Another crucial issue is the quality of experimental samples, which is often not sufficiently known or characterized. In fact, the sample's quality depends on its history. Metadata about the experiment, the sample, and the data must be defined if the experimental data is reused or combined with other data. This is rarely done currently. The NOMAD team, together with experimental colleagues,[49] is taking the first steps in this direction.

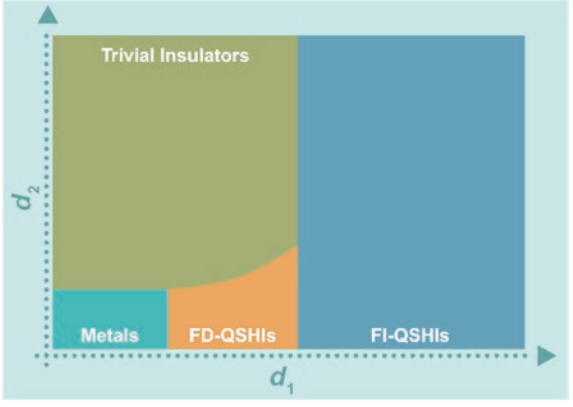


Figure 5. A map of materials classifying 2D functionalized honeycomb structures as functionalization-independent (FI) quantum spin hall insulators (QSHIs), functionalization-dependent (FD) QSHIs, metals, and trivial insulators. Note: d_1 and d_2 are the descriptors identified in Reference 48.

https://doi.org/10.1557/mrs.2018.208 Published online by Cambridge University Press

Acknowledgments

This work received funding from the EU's Horizon 2020 Research and Innovation Programme, Grant Agreement No. 676580, the NOMAD Laboratory CoE and No. 740233, ERC: TEC1P. Support from the Einstein Foundation Berlin is appreciated. We thank C. Koch for in-depth discussions and information about transmission electron microscopy, P. Wittenburg for clarification of the FAIR concept, and the whole NOMAD team for the invaluable effort to build the entire NOMAD infrastructure and its services.

*We note that accessibility in this context not only concerns the data itself, but largely also the hardware at which it is stored, as well as details of the data infrastructure.

†The concept of the NOMAD Repository and Archive was developed in 2014, independently and parallel to the "FAIR Guiding Principles."[10] Interestingly, the substance is practically identical.

‡The understanding of "interoperable" is somewhat controversial. It may apply when (1) metadata and data use a formal, accessible, shared, and widely accepted language for data representation; (2) the vocabulary of metadata and data follow FAIR principles; (3) metadata and data include qualified references to other (meta) data and to the authors who created the results.

§In the NOMAD CoE, we use the term repurposable, while in the FAIR concept it was termed reusable. Both obviously mean the same in this context.

**The concept of a fourth paradigm was probably first discussed by J. Gray on January 11, 2007, before he went missing in the Pacific Ocean on January 28, 2007.[29]

††"Codes" is widely used jargon in the field for scientific software packages.

‡‡The Psi-k network represents scientists concerned with *ab initio* calculations of materials.

§§CECAM is a European organization devoted to the promotion of fundamental research on advanced computational methods and to their application to important problems in frontier areas of science and technology, in physics and chemistry of condensed matter.

***ETSF, the European Theoretical Spectroscopy Facility, is a knowledge center for theoretical spectroscopy.

References

1. S. Curtarolo, G.L.W. Hart, M.B. Nardelli, N. Mingo, S. Sanvito, O. Levy, *Nat. Mater.* **12**, 191 (2013).
2. J.E. Saal, S. Kirklin, M. Aykol, B. Meredig, C. Wolverton, *JOM* **65**, 1501 (2013).
3. I.E. Castelli, T. Olsen, S. Datta, D.D. Landis, S. Dahl, K.S. Thygesen, K.W. Jacobsen, *Energy Environ. Sci.* **5**, 5814 (2012).
4. M. Nishijima, T. Ootani, Y. Kamimura, T. Sueki, S. Esaki, S. Murai, K. Fujita, K. Tanaka, K. Ohira, Y. Koyama, I. Tanaka, *Nat. Commun.* **5**, ncomms5553 (2014).
5. J. Hachmann, R. Olivares-Amaya, S. Atahan-Evrenk, C. Amador-Bedolla, R.S. Sánchez-Carrera, A. Gold-Parker, L. Vogt, A.M. Brockway, A. Aspuru-Guzik, *J. Phys. Chem. Lett.* **2**, 2241 (2011).
6. https://www.mgi.gov.
7. J.A. Warren, R.F. Boisvert, "Workshop Report: Building the Materials Innovation Infrastructure: Data and Standards A Materials Genome Initiative Workshop" NIST Report No. NISTIR 7898, (2012).
8. "Empty Rhetoric over Data Sharing Slows Science" [Editorial], *Nature* **546**, 327 (2017).
9. The NOMAD (Novel Materials Discovery) Center of Excellence (CoE), https://nomad-coe.eu.
10. M.D. Wilkinson, M. Dumontier, I.J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L.B. da Silva Santos, P.E. Bourne, J. Bouwman, A.J. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C.T. Evelo, R. Finkers, A. Gonzalez-Beltran, A.J.G. Gray, P. Groth, C. Goble, J.S. Grethe, J. Heringa, P.A.C. 't Hoen, R. Hooft, T. Kuhn, R. Kok, J. Kok, S.J. Lusher, M.E. Martone, A. Mons, A.L. Packer, B. Persson, P. Rocca-Serra, M. Roos, R. van Schaik, S.-A. Sansone, E. Schultes, T. Sengstag, T. Slater, G. Strawn, M.A. Swertz, M. Thompson, J. van der Lei, E. van Mulligen, J. Velterop, A. Waagmeester, P. Wittenburg, K. Wolstencroft, J. Zhao, B. Mons, *Sci. Data* **3**, 160018 (2016).
11. A. Jain, S.P. Ong, G. Hautier, W. Chen, W.D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, K.A. Persson, *APL Mater.* **1**, 011002 (2013).
12. L.M. Ghiringhelli, C. Carbogno, S. Levchenko, F. Mohamed, G. Huhs, M. Lueders, M. Oliveira, M. Scheffler, *Psi-k Scientific Highlight of the Month* **131** (2016).
13. L.M. Ghiringhelli, C. Carbogno, S. Levchenko, F. Mohamed, G. Huhs, M. Lueders, M. Oliveira, M. Scheffler, *NPJ Comput. Mater.* **3**, 46 (2017).
14. The NOMAD Archive, https://metainfo.nomad-coe.eu/nomadmetainfo_public/archive.html.
15. K. Rajan, *Mater. Today* **8**, 35 (2005).
16. M.W. Gaultois, A.O. Oliynyk, A. Mar, T.D. Sparks, G.J. Mulholland, B. Meredig, *APL Mater.* **4**, 053213 (2016).
17. T. Mueller, A.G. Kusne, R. Ramprasad, in *Reviews in Computational Chemistry*, A.L. Parrill, K.B. Lipkowitz, Eds. (Wiley, Hoboken, NJ), vol. 29, p. 186.
18. T. Mueller, E. Johlin, J.C. Grossman, *Phys. Rev. B Condens. Matter* **89**, 115202 (2014).
19. R. Ramprasad, R. Batra, G. Pilania, A. Mannodi-Kanakkithodi, C. Kim, *NPJ Comput. Mater.* **3**, 54 (2017).
20. G. Pilania, C.C. Wang, X. Jiang, S. Rajasekaran, R. Ramprasad. *Sci. Rep.* **3**, 2810 (2013).
21. A.V. Lilienfeld, *Angew. Chem. Int. Ed. Engl.* **57**, 4164 (2018).
22. A. Ziletti, D. Kumar, M. Scheffler, L.M. Ghiringhelli, *Nat. Commun.* **9**, 2775 (2018).
23. Q. Yang, C.A. Sing-Long, E.J. Reed, *Chem. Sci.* **8**, 5781 (2017).
24. L.M. Ghiringhelli, J. Vybiral, S.V. Levchenko, C. Draxl, M. Scheffler, *Phys. Rev. Lett.* **114**, 105503 (2015).
25. L.M. Ghiringhelli, J. Vybiral, E. Ahmetcik, R. Ouyang, S.V. Levchenko, C. Draxl, M. Scheffler, *New J. Phys.* **19**, 023017 (2017).
26. R. Ouyang, S. Curtarolo, E. Ahmetcik, M. Scheffler, L.M. Ghiringhelli, *Phys. Rev. Mater.* **2**, 083802 (2018).
27. K. Hansen, G. Montavon, F. Biegler, S. Fazli, M. Rupp, M. Scheffler, O.A. von Lilienfeld, A. Tkatchenko, K.-R. Müller, *J. Chem. Theory Comput.* **9**, 3404 (2013).
28. B.R. Goldsmith, M. Boley, J. Vreeken, M. Scheffler, L.M. Ghiringhelli, *New J. Phys.* **19**, 013031 (2017).
29. J. Gray, *The Fourth Paradigm, Data Intensive Discovery*, T. Hey, S. Tansley, K. Tolle, Eds. (Microsoft Research, Redmond, WA, 2009).
30. J.A. Van Vechten, *Phys. Rev.* **182**, 891 (1969).
31. J.C. Phillips, *Rev. Mod. Phys.* **42**, 317 (1970).
32. M. Ashby, *Materials Selection in Mechanical Design*, 3rd ed. (Butterworth-Heinemann, Burlington, MA, 1999).
33. K. Lejaeghere, G. Bihlmayer, T. Björkman, P. Blaha, S. Blügel, V. Blum, D. Caliste, I.E. Castelli, S.J. Clark, A. Dal Corso, S. de Gironcoli, T. Deutsch, J.K. Dewhurst, I. Di Marco, C. Draxl, M. Dułak, O. Eriksson, J.A. Flores-Livas, K.F. Garrity, L. Genovese, P. Giannozzi, M. Giantomassi, S. Goedecker, X. Gonze, O. Grånäs, E.K.U. Gross, A. Gulans, F. Gygi, D.R. Hamann, P.J. Hasnip, N.A.W. Holzwarth, D. Iuşan, D.B. Jochym, F. Jollet, D. Jones, G. Kresse, K. Koepernik, E. Küçükbenli, Y.O. Kvashnin, I.L.M. Locht, S. Lubeck, M. Marsman, N. Marzari, U. Nitzsche, L. Nordström, T. Ozaki, L. Paulatto, C.J. Pickard, W. Poelmans, M.I.J. Probert, K. Refson, M. Richter, G.M. Rignanese, S. Saha, M. Scheffler, M. Schlipf, K. Schwarz, S. Sharma, F. Tavazza, P. Thunström, A. Tkatchenko, M. Torrent, D. Vanderbilt, M.J. van Setten, V. Van Speybroeck, J.M. Wills, J.R. Yates, G.X. Zhang, S. Cottenier, *Science* **351**, aad3000 (2016).
34. The Novel Materials Discovery (NOMAD) Repository, https://repository.nomad-coe.eu.
35. The NOMAD Encyclopedia, https://encyclopedia.nomad-coe.eu.
36. http://www.psi-k.org.
37. https://www.cecam.org.
38. https://www.etsf.eu.
39. http://www.aflowlib.org.
40. http://oqmd.org.
41. https://materialsproject.org.
42. *Scientific Data*, a journal of the Nature Publishing Group, https://www.nature.com/sdata.
43. C. Draxl, F. Illas, M. Scheffler, *Nature* **548**, 523 (2017).
44. NOMAD—The Materials Science Discovery Repository, https://youtu.be/UcnHGokl2Nc, 2017-05-30.
45. Conversion of CO_2 into Fuels and Other Useful Chemicals, https://youtu.be/zHIS_8PwYYs, 2017-06-23.
46. An Exciton in Lithium Fluoride—Where Is the Electron? https://youtu.be/XPPDeeP1coM, 2017-06-23.
47. The NOMAD Analytics Toolkit, https://analytics-toolkit.nomad-coe.eu/home.
48. C. Mera Acosta, R. Ouyang, A. Fazzio, M. Scheffler, L.M. Ghiringhelli, C. Carbogno, arXiv:1805.10950 (2018).
49. https://www.bigmax.mpg.de. □

https://doi.org/10.1557/mrs.2018.208 Published online by Cambridge University Press



Claudia Draxl is a full professor at the Humboldt-Universität and Max Planck Fellow at the Fritz-Haber-Institut Berlin, Germany. She studied physics and mathematics at the University of Graz, Austria, where she received her PhD degree in solid-state theory. Her current research focuses on the understanding of conventional and organic semiconductors and interfaces thereof, complex alloys, and two-dimensional systems. A major focus is excited states, using and developing techniques beyond density functional theory. She has authored approximately 250 publications. She received the Ludwig Boltzmann Award of the Austrian Physical Society and was conferred an honorary doctorate by Uppsala University, Sweden. Draxl can be reached by email at claudia.draxl@physik.hu-berlin.de.



Matthias Scheffler is director at the Fritz-Haber-Institut Berlin of the Max Planck Society, Germany. He is also an honorary professor of theoretical condensed-matter physics at all three Berlin universities and a distinguished visiting professor for materials science and engineering at the University of California, Santa Barbara, and Hokkaido University (Sapporo), Japan. His research focuses on the basic aspects of physical and chemical properties and processes at bulk materials, surfaces, interfaces, defects, clusters, and nanostructures. He developed various methods for first-principles computations of the electronic structure and systematic hybrid approaches of electronic-structure theory, thermodynamics, and statistical mechanics. His recent studies focus on big data-driven materials science. Scheffler can be reached by email at scheffler@fhi-berlin.mpg.de.

CALL FOR PAPERS

jmr Journal of MATERIALS RESEARCH



MAY 2019 Plasticity and Fracture at the Nanoscales–Advances in *In Situ* Experimentation Techniques Enabling Novel and Extreme Materials/Nanocomposite Design
Submission Deadline: October 1, 2018



JUNE 2019
(Nano)materials for Biomedical Applications
Submission Deadline: November 1, 2018



JULY 2019
Intrinsic and Extrinsic Size Effects in Materials
Submission Deadline: December 1, 2018

www.mrs.org/jmr

https://doi.org/10.1557/mrs.2018.208 Published online by Cambridge University Press