ESSAY

Databases

ADVANCED
THEORY AND
SIMULATIONS

www.advtheorysimul.com

Materials Databases: The Need for Open, Interoperable Databases with Standardized Data and Rich Metadata

François-Xavier Coudert

Driven by the recent rapid increase in the number of materials databases published (open and commercial), some perspectives on the growing need for standardized, interoperable, open databases are discussed here. The field of computational materials discovery is quickly expanding, and recent advances in data mining, high-throughput screening, and machine learning highlight the potential of open databases.

1. Introduction

One of the recent important trends in materials science is the emergence of several large-scale online databases of materials, trying to bring together experimental data with computational techniques in order to understand the behavior of materials families and design novel materials through data-mining? Maybe the most visible efforts in this area are the Materials Project,[1] a database of computed information on known and predicted properties, part of the US-funded Materials Genome Initiative[2] launched in 2011; and the very similar AFLOW initiative, a distributed materials properties repository from high-throughput ab initio calculations.[3] However, this trend is more general, and an increasing amount of research in the field focuses on the generation of these databases, their extension with additional data, and the use of these databases for analysis, screening, and prediction. This was clearly exemplified at recent materials meetings and materials modeling workshop, such as the MOFSIM 2019[4] meeting whose excellent discussion provoked this short comment.

2. Existing Databases

The need for aggregation of curated data in the physical and chemical sciences was recognized very early, and possibly the best-known database in our field is the CRC Handbook of Chemistry and Physics[5] (also known as the "rubber book"), which has been published since 1914. When it comes to materials,

databases ranging in size and scope have emerged since the beginning of computing, and have advanced at the same pace as both computer hardware and networking capabilities. The Cambridge Structural Database (CSD),[6] launched in 1965, is one of the first numerical scientific databases, operating as a repository for validated experimental crystal structures of organic and organometallic compounds. Additional structural databases have emerged focusing on other categories of materials, including the Inorganic Crystal Structure Database (ICSD),[7] the Protein Data Bank (PDB),[8] the American Mineralogist Crystal Structure Database,[9] and the Crystallography Open Database (COD).[10]

In addition to these structural databases, databases of materials properties have also been compiled, either by standardization institutes, learned societies, or commercial entities. We can cite examples of the NIST databases for materials and fluids properties, the glass property database SciGlass, Pearson's Crystal Data, or more specific examples like the Polymer Gas Separation Membrane Database from the Membrane Society of Australasia. Moreover, databases of hypothetical structures—predicted by theory or computations—have also appeared over time. In the field of porous materials, for example, databases of computed zeolitic structures have been published more than 15 years ago.[11,12] More recently, this effort has intensified—probably in response to the increase in capacity of computations, as well as the ease of hosting large datasets online. Databases of various scales, containing hypothetical (enumerated) metal–organic frameworks, have been published.[13] Other groups have worked on refining experimental databases to include additional data (such as atomic charges), making them suitable for computational applications and screening.[14,15] All these databases, however, are hosted independently as archive files, with heterogeneous file formats, on individual research groups' websites.

We note, however, that there have been some recent initiatives in order to integrate data from different sources into larger, coherent databases. This is particularly the case of computational data, whose volume increases with high-performance computing capabilities. The goals differ for the various initiatives, but in general, they aim at providing large-scale platforms for open science and data sharing, as well as improve discoverability and searchability of existing data. A first example is the Materials Project,[1] the aim of which is to "remove guesswork from materials design in a variety of applications" by computing properties of all known materials (and many that are not yet synthesized, too) through electronic structure analyses—a project funded as part of the bigger Materials Genome Initiative.[2] It aggregates

Dr. F.-X. Coudert
Chimie ParisTech
PSL University
CNRS, Institut de Recherche de Chimie Paris
75005 Paris, France
E-mail: fx.coudert@chimieparistech.psl.eu

The ORCID identification number(s) for the author(s) of this article can be found under https://doi.org/10.1002/adts.201900131

ADVANCED
SCIENCE NEWS
www.advancedsciencenews.com

ADVANCED
THEORY AND
SIMULATIONS
www.advtheorysimul.com

structural data from other existing databases, as well as physical and chemical properties (band structures, elastic constants, piezoelectric tensors, electrode properties) computed as part of the project itself. Other platforms have been built to sharing results and resources in computational materials sciences, such as AFLOW repository,[3] the NOMAD repository,[16,17] the Open Quantum Materials Database,[18] the NRELMatDB database (focused on materials for renewable energy applications).[19] We can also cite here the ioChem-BD digital repository[20] and the more recent Materials Cloud.[21]

3. The Current State of Affairs

As shown above, the number of existing databases is increasing at a fast pace. Yet, the datasets themselves are often hosted using makeshift solutions, in different places. They are typically too large to be hosted as supporting information by the publisher of the associated research paper, and they might be expanded and refined over time, with the publication of updates not related to a particular peer-reviewed paper. In the most common case, they are hosted on the web server of the research group, on institutional repositories offered by some universities, or on free data hosting solutions such as Figshare,[22] GitHub,[23] etc. This leads to the available data being dispersed between several platforms, and in some cases, raises the question of long-term availability of the data (e.g., when it is hosted on a group web server or a commercial data hosting solution). Moreover, there exists no universal way to access these data, unlike what exists for open archives where protocols such as OAI-PMH have been developed for interoperability and discoverability of data sources.

This fragmentation of the landscape of materials databases is accompanied by a large heterogeneity in the formats used: while for crystalline material structures, the CIF (crystallographic information file) format is predominant, the data made available as part of the CIF file is not homogeneous between different groups. The use of symmetry operators, for example, is not always consistent, with some databases being stored with symmetry systematically lowered to $P1$. In addition to this heterogeneity in data format, there is also a general lack of availability of metadata, meaning that most of the databases do not contain information about how the data was generated, gathered, curated, and possibly updated. Yet, this metadata can be crucial in exploiting databases, in order to identify identical or related data items, to understand how databases evolve over time, and to allow further investigation of specific data items. Metadata enables researchers to answer simple queries such as: where does this structure come from? Where was it first reported, and under what conditions was it synthesized? How was this computational property calculated? What are the conditions of reuse of this data?

Here, before highlighting some requirements for truly open materials databases, we want to introduce the FAIR data principles. The FAIR principles are a set of guidelines in order to make data findable, accessible, interoperable, and reusable (**Figure 1**). They have been formalized in 2016 by a consortium of scientists and organizations,[24] and were formally endorsed by the G20 Leaders at their 2016 summit in Hangzhou (杭州), China, in order to "promote open science and facilitate appropriate access

François-Xavier Coudert is a Senior Researcher at the French CNRS. His group at PSL University applies computational chemistry methods at various scales to investigate the physical and chemical properties of nanoporous materials, and in particular stimuli-responsive materials with anomalous behavior.

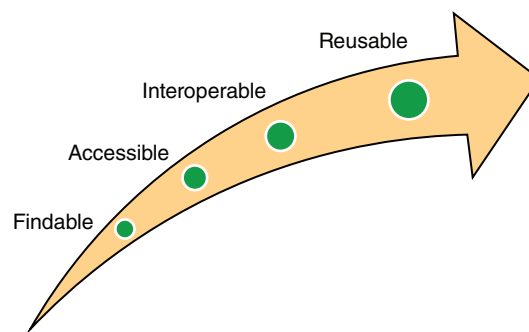to publicly funded research results".[25] Requirements for FAIR data are:

- **Findable**: data are indexed in a searchable resource, has a stable unique identifier, and is described with rich metadata.
- **Accessible**: data are retrievable using a standardized communications protocol, that is open, free, and universally implementable.
- **Interoperable**: data are represented using an open, well-defined format; data and metadata are interlinked.
- **Reusable**: data contains relevant metadata about its origin, a clear and accessible data usage license, and meet the community standards of its domain

4. Requirements For Open Databases

Based on these formal requirements, and during the discussions at MOFSIM 2019 and other workshops, there appears to be a need for more open and interoperable materials databases. We outline below some of the requirements, drawing both on the FAIR principles and shared experiences and discussions.

4.1. Open Databases

A majority of the available data on materials is produced as part of academic research, and largely funded by public monies. Moreover, for data associated with public research, there is a general

Figure 1. FAIR data principles[24]: make data findable, accessible, interoperable, and reusable.

ADVANCED
SCIENCE NEWS
www.advancedsciencenews.com

ADVANCED
THEORY AND
SIMULATIONS
www.advtheorysimul.com

consensus that it should be accessible to all. Recent years have seen the addition of "data availability" requirements in several journals, and it is now a required part for any funding application. This creates a need for open databases, where content can be hosted regardless of its origin (contrary to institution-wide repositories) and where it is accessible to all. Moreover, the database should provide clear information about its users' rights when it comes to reusing the data, mining it, and republishing derived products.

Because open databases do not charge their users for subscription or access, they allow the dissemination of knowledge to categories of users that would otherwise find it difficult to obtain access: researchers in developing countries, nongovernmental organizations, independent researchers, journalists, even enthusiastic citizens. Operating such open databases of course requires funding, and it is important to note that a number of national and supranational initiatives have been launched in that direction (as discussed in the introduction). This is an important requirement for the availability of high-quality open databases, and ought to be considered at the same time as shifting priorities for open access to publication. For example, the recent French National Plan for Open Science[26] has lead to specific calls for funding for open data, and several field-specific databases have been selected for funding through this scheme.

4.2. Interoperable Databases

It is relatively clear that, given the vastly different needs of scientists working in different areas of materials science, there can be no "one size fits all" database, that is, no single centralized database that fulfills the needs of every different community. So, how can a good balance be reached in developing specific topical databases while retaining some uniformity, in order to avoid a complete fragmentation of the field? It turns out, this problem is one that has been worked on for many years in a related area, that of document (or papers) archives. While there are many different open archives on the internet, they have been developed in a way that allows interoperability between them. Specifically, the Open Archives Initiative (OAI) has standardized a Protocol for Metadata Harvesting (OAI-PMH) through which each archive exposes its metadata, in a common format, allowing for cross-database search and discoverability.[27]

In order to achieve this goal, several design choices are needed. One is the use of a well-documented, standardized Application Programming Interface (API). Through the use of that API, the data does not have to be retrieved with a database-specific client or web portal,[28] but can be written in any programming language without inside knowledge of how the database operates internally. This means, in turn, that code that is developed for one specific database will work seamlessly with all others.

Another is the inclusion of data in standard, publicly documented file formats. Given that most current databases are currently structural databases, a part of this problem has already been addressed in the several past decades: crystal structures are uniformly reported in CIF format (although the details available are not always consistent), macromolecular structures are consistently in PDB format, etc. However, there is currently no unified format for storing the properties of these materials.

This is made difficult by the fact that properties are rather diverse in their mathematical nature: some are dimensionless but others have units; some are integers or half-integers, others vary continuously; some are scalars, others are matrices or higher-order tensors. Moreover, they sometimes need to be accompanied by additional information: expected uncertainty, reference orientation, etc.

4.3. With Rich Metadata and Interlinked Datasets

Metadata can be defined, in its simplest form, as "data that provides information about other data." Like in any database, in a materials database, metadata is crucial in assessing the data present, answering questions such as: how was this data gathered, by whom, when? In which conditions was a given property measured? For computational information, what was the theoretical method used, what is the level of description of the system? This is particularly important in databases of computational properties, where there can be a clear influence—and sometimes even a systematic bias—of the computational method chosen on the physical and chemical data calculated. If metadata is present in the databases, it opens the door to large-scale systematic explorations of various theoretical methods, and their comparison with experimental results obtained with different techniques, too.

Moreover, metadata can also provide much-needed links between several different interoperable databases. If a unique identifier is given for each dataset, and databases are interlinked through their metadata, it provides a much simpler exploration for users. It makes it easy to determine, for example, if two properties from two datasets are independent or come from the same original calculation. It also allows greater discoverability, making it possible to find other properties in other databases, related to any given entry.

4.4. Curation That Preserves the Scientific Record

The requirements listed above do not stop a fixed point of time, but instead must be considered throughout the databases' timeline. For example, metadata can record the time of measurement of a given data, but also its time of inclusion in the database, and its further history. Indeed, with any database of significant size, it is expected that curation of the data is an important topic, and the dataset will be modified to remove errors, updated to reflect new measurements, and sometimes data will be removed for a variety of legitimate reasons. However, for the sake of research reproducibility and conserving the scientific record, it is important that such modifications be recorded in the metadata—just like corrections and retractions are publicly announced and archived for scientific articles. To my knowledge, this is not currently the case in existing databases, although the Materials Project is publishing "release logs"[29] which are kept on a separate page, but not recorded in the database as metadata.

4.5. Long-Term Availability

Finally, this discussion cannot be concluded without addressing the issue of long-term availability of the deposited data, meaning

**ADVANCED
SCIENCE NEWS**

www.advancedsciencenews.com

**ADVANCED
THEORY AND
SIMULATIONS**

www.advtheorysimul.com

that it is necessary, over time, to build institutional support with long-term commitments. This also requires planning for what happens if and when the hosting institutions decide to "pull the plug" on the project. There, having an open database with an API for direct access to bulk data is a benefit, because it means other interested parties can duplicate the database and take over hosting. Other lessons can be learned from open archives, and initiatives in that field such as CLOCKSS, a long-term preservation project for articles and books with highly redundant mirroring.[30]

4.6. Validation, Interlinking with Experimental Data

The computational databases discussed above are almost exclusively composed of calculated materials properties, based on either experimental or hypothetical structures. These properties are sometimes benchmarked against experimental results in publications—especially for the initial validation of the methodology—but it is rarely done in a systematic manner, and integrated into the database itself. Now that large-scale computational databases are available, and are starting to appear at different levels of theory, there is a need for systematic validation of large subsets against experimental data, and inclusion (or interlinking) of this experimental data in a systematic manner with the computational databases.

5. Application to Metal–Organic Frameworks

The computational databases listed above have mainly focused on properties of materials with relatively simple chemical composition and structure, and in particular inorganic crystalline compounds—they form the vast majority of the datasets available today. However, given the rapidly growing number of metal–organic frameworks (MOFs) reported and present in licensed databases like the Cambridge Structural Database (CSD), there have been initiatives in this area as well. However, there have been until recently no real computational databases with reported materials properties, but research instead focused mostly on structural databases—whether from experimental data or hypothetical frameworks. The former exploited the CSD database to apply filters and high-throughput screening of, for example, adsorption properties: this was, for example, the case of Goldsmith et al. for a study on hydrogen storage.[31] The later used combinatorial algorithms to identify novel hypothetical MOF, such as the seminal work of Wilmer et al. to obtain a database of 137 953 hypothetical MOF structures from 102 different building blocks.[13]

Beyond this initial work, more recent studies have shifted to the production of so-called "computationally ready" structural databases: these are curated databases of structures adapted from experimental crystallographic files, so that they are suitable for running calculations in a routine (and high-throughput compatible) way. This involves the automated removal of included solvent molecules and disorder, addition of missing hydrogens, and general cleaning up of the structures. The first such published database was the CoRE MOF database,[14] containing more than 6000 3D MOF structures. Another example is the recent CSD filter by Moghadam et al.,[32] built on top of the commercial CSD

database—and which thus stays up-to-date when new structures are added.

Finally, I note that some MOF structural databases have been augmented, over time, with additional computational data: for example, Nazarian et al. added to a subset of the Core MOF database atomic point charges derived from periodic DFT electronic structure calculations.[15,33] There is, however, to this date no existing database of MOFs with macroscopic materials properties (vibration modes, elastic properties, etc.)—this is probably due to the much larger computational effort required for such calculations, compared to "simpler" inorganic compounds. We may thus expect that situation to change in the future.

Conflict of Interest

The author declares no conflict of interest.

Keywords

databases, metadata, open science

[1] 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.* **2013**, *1*, 011002.

[2] A. White, *MRS Bull.* **2012**, *37*, 715.

[3] S. Curtarolo, W. Setyawan, S. Wang, J. Xue, K. Yang, R. H. Taylor, L. J. Nelson, G. L. Hart, S. Sanvito, M. Buongiorno-Nardelli, N. Mingo, O. Levy, *Comp. Mater. Sci.* **2012**, *58*, 227.

[4] MOFSIM2019 Meeting, Ghent, Belgium, April 10–12, **2019**, https://mofsim.eu.

[5] J. R. Rumble, *CRC Handbook of Chemistry and Physics*, 99th ed., CRC Press, Boca Raton, FL **2018**.

[6] C. R. Groom, I. J. Bruno, M. P. Lightfoot, S. C. Ward, *Acta Cryst. B* **2016**, *72*, 171.

[7] M. Hellenbrandt, *Cryst. Rev.* **2014**, *10*, 17.

[8] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, P. E. Bourne, *Nucleic Acids Res.* **2000**, *28*, 235.

[9] R. T. Downs, M. Hall-Wallace, *Am. Mineral.* **2003**, *88*, 247.

[10] S. Gražulis, A. Daškevič, A. Merkys, D. Chateigner, L. Lutterotti, M. Quirós, N. R. Serebryanaya, P. Moeck, R. T. Downs, A. Le Bail, *Nucleic Acids Res.* **2012**, *40*, D420.

[11] Y. Li, J. Yu, D. Liu, W. Yan, R. Xu, Y. Xu, *Chem. Mater.* **2003**, *15*, 2780.

[12] D. J. Earl, M. W. Deem, *Ind. Eng. Chem. Res.* **2006**, *45*, 5449.

[13] C. E. Wilmer, M. Leaf, C. Y. Lee, O. K. Farha, B. G. Hauser, J. T. Hupp, R. Q. Snurr, *Nature Chem.* **2012**, *4*, 83.

[14] Y. G. Chung, J. Camp, M. Haranczyk, B. J. Sikora, W. Bury, V. Krungleviciute, T. Yildirim, O. K. Farha, D. S. Sholl, R. Q. Snurr, *Chem. Mater.* **2014**, *26*, 6185.

[15] D. Nazarian, J. S. Camp, D. S. Sholl, *Chem. Mater.* **2016**, *28*, 785.

[16] L. M. Ghiringhelli, C. Carbogno, S. Levchenko, F. Mohamed, G. Huhs, M. Lüders, M. Oliveira, M. Scheffler, *npj Comput. Mater.* **2017**, *3*, 190.

[17] C. Draxl, M. Scheffler, *MRS Bull.* **2018**, *43*, 676.

**ADVANCED
SCIENCE NEWS**

www.advancedsciencenews.com

**ADVANCED
THEORY AND
SIMULATIONS**

www.advtheorysimul.com

[18] S. Kirklin, J. E. Saal, B. Meredig, A. Thompson, J. W. Doak, M. Aykol, S. Rühl, C. Wolverton, *npj Comput. Mater.* **2015**, *1*, 864.

[19] NREL MatDB home, http://materials.nrel.gov

[20] ioChem-BD, home page, https://www.iochem-bd.org/

[21] Materials Cloud, home page, https://www.materialscloud.org/

[22] GitHub, home page, https://www.github.com/

[23] figshare, home page, https://www.figshare.com/

[24] 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. Gray, P. Groth, C. Goble, J. S. Grethe, J. Heringa, P. A. 't Hoen, R. Hooft, T. Kuhn, R. Kok, J. Kok, et al., *Sci. Data* **2016**, *3*, e1002295.

[25] G20 Leaders' Communiqué, Hangzhou Summit, Hangzhou, China, September 5, **2016**, http://europa.eu/rapid/press-release_STATEMENT-16-2967_en.htm

[26] https://www.ouvrirlascience.fr/open-science/

[27] The Open Archives initiative protocol for metadata harvesting, http://www.openarchives.org/OAI/openarchivesprotocol.html

[28] Although of course such clients can exist, providing a user-friendly way to query the database! The existence of a public API makes it possible for advanced users to develop their own portals, bringing added value to the database.

[29] Materials project database release log, https://discuss.materialsproject.org/t/materials-project-database-release-log/1609

[30] CLOCKSS, home page, https://clockss.org

[31] J. Goldsmith, A. G. Wong-Foy, M. J. Cafarella, D. J. Siegel, *Chem. Mater.* **2013**, *25*, 3373.

[32] P. Z. Moghadam, A. Li, S. B. Wiggin, A. Tao, A. G. P. Maloney, P. A. Wood, S. C. Ward, D. Fairen-Jimenez, *Chem. Mater.* **2017**, *29*, 2618.

[33] D. Nazarian, J. S. Camp, Y. G. Chung, R. Q. Snurr, D. S. Sholl, *Chem. Mater.* **2016**, *29*, 2521.

1900131 (5 of 5)