scientific data

OPEN

COMMENT

Shared metadata for data-centric materials science

Luca M. Ghiringhelli *et al.*#

The expansive production of data in materials science, their widespread sharing and repurposing requires educated support and stewardship. In order to ensure that this need helps rather than hinders scientific work, the implementation of the FAIR-data principles (*Findable*, *Accessible*, *Interoperable*, *and Reusable*) must not be too narrow. Besides, the wider materials-science community ought to agree on the strategies to tackle the challenges that are specific to its data, both from computations and experiments. In this paper, we present the result of the discussions held at the workshop on "Shared Metadata and Data Formats for Big-Data Driven Materials Science". We start from an operative definition of metadata, and the features that a FAIR-compliant metadata schema should have. We will mainly focus on computational materials-science data and propose a constructive approach for the *FAIRification* of the (meta)data related to ground-state and excited-states calculations, potential-energy sampling, and generalized workflows. Finally, challenges with the *FAIRification* of experimental (meta)data and materials-science ontologies are presented together with an outlook of how to meet them.

Introduction: Metadata and FAIR data principles

The amount of data that has been produced in materials science till today and its day-by-day increase are massive[1]. The dawn of the data-centric era[2] requires that such data are not just stored, but also carefully annotated in order to find, access, and possibly reuse them. Terms of good practice to be adopted by the scientific community for the management and stewardship of its data, the so-called FAIR-data principles, have been compiled by the FORCE11 group[3]. Here, the acronym FAIR stands for *Findable, Accessible, Interoperable, and Reusable*, which applies not only to *data*, but also to *metadata*. Other terms for the "R" in FAIR are "repurposable" and "recyclable". The former term indicates that data may be used for a different purpose than the original one for which they were created. The latter term hints at the fact that data in materials science are often exploited only once for supporting the thesis of a single publication and then they are stored and forgotten. In this sense, they would constitute a "waste" that can be recycled, provided that they can be found and they are properly annotated.

Before examining the meaning and importance of the four terms of the FAIR acronym, it is worth defining what metadata are with respect to data. To the purpose, we start by introducing the concept of *data object*. A *data object* is the collective storage of information related to an elementary entry in a database. One can consider it as a row in a table, where the columns can be occupied by simple scalars, higher-order mathematical objects, strings of characters, or even full documents (or other media objects). In the materials-science context, a *data object* is the collection of attributes (the columns in the above-mentioned table) that represent a material or, even more fundamentally, a *snapshot* of the material captured by a single configuration of atoms, or it may be a set of measurements from well-defined *equivalent samples* (see below for a discussion on this concept). For instance, in computational materials science, the attributes of a *data object* could be both the inputs (e.g., the coordinates and chemical species of the atoms constituting the material, the description of the physical model used for calculating its properties), and the outputs (e.g., total energy, forces, electronic density of states, etc.) of a calculation. Logically and physically, inputs and outputs are at different levels, in the sense that the former determine the latter. Hence, one can consider the inputs as *metadata* describing the *data*, i.e., the outputs. In turn, the set of coordinates A that are metadata to some observed quantities, may be considered as data that depend on another set of coordinates B, and the forces acting on the atoms in that set A. So, the set of coordinates B and the acting forces are metadata to the set A, now regarded as data. Metadata can always be considered to be data as they could be objects of different, independent analyses than those performed on the calculated properties. In

#A full list of authors and their affiliations appears at the end of the paper.

this respect, whether an attribute of a *data object* is data or metadata depends on the *context*. This simple example also depicts a *provenance* relationship between the data and their metadata.

The above discussion can be summarized in a more general definition of the term metadata: *Metadata are attributes that are necessary to locate, fully characterize, and ultimately reproduce other attributes that are identified as data*. The metadata include a clear and unambiguous description of the data as well as their full provenance. This definition is reminiscent of the definition given by NIST[4]: "Structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use, or manage an information resource. Metadata is often called data about information or information about information". With our definition, we highlight the role of data "reproducibility", which is crucial in science.

Within the "full characterization" requirement, we highlight *interpretation* of the data as a crucial aspect. In other words, the metadata must provide enough information on a stored value (therein including, e.g., adimensional constants) to make it unambiguous whether two data objects may be compared with respect to the value of a given attribute or not.

Next, we should notice that, whereas in computational materials science the concept of data object identified by a single atomic configuration is well defined, in experimental materials science the concept of a class of *equivalent samples* is very hard to implement operationally. For instance, a single specimen can be altered by a measurement operation and thus cannot, strictly speaking, be measured twice. At the same time, two specimens prepared with the same synthesis recipe, may differ in substantial aspects due to the presence of different impurities or even crystal phases, thus yielding different values of a measured quantity. In this respect, here we use the term *equivalent sample* in its abstract, ideal meaning, but we also mention that one of the main purposes of introducing well-defined metadata in materials science is to provide enough characterization of experimental samples to put into practice the concept of *equivalent samples*.

The need for storing and characterizing data by means of metadata is determined by two main aspects, related to data usage. The first aspect is as old as science: *reproducibility*. In an experiment or computation, all the necessary information needed to reproduce the measured/calculated data (i.e., the metadata) should be recorded, stored, and retrievable. The second aspect becomes prominent with the demand for *reusability*. Data can and should be also usable for purposes that were not anticipated at the time they were recorded. A useful way of looking at metadata is that they are attributes of *data objects* answering the "wh- questions": who, what, when, where, why, and how. For example, "Who has produced the data?", "What are the data expected to represent (in physical terms)?", "When were they produced?", "Where are they stored?", "For what purpose were they produced?", and "By means of which methods were the data obtained?". The latter two questions also refer to the concept of *provenance*, i.e., the logical sequence of operations that determine, ideally univocally, the data. Keeping track of the provenance requires the possibility to record the whole *workflow* that has lead to some calculated or measured properties (for more details, see Section "Metadata for Computational Workflows").

From a practical point of view, the metadata are organized in a schema. We summarize what the FAIR principles imply in terms of a metadata schema as follows:

- *Findability* is achieved by assigning unique and Persistent Identifiers (PIDs) to data and metadata, describing data with rich metadata, and *registering* (see below) the (meta)data in searchable resources. Widely known examples of PIDs are digital object identifiers (DOIs) and (permanent) Uniform Resource Identifiers (URIs). According to ISO/IEC 11179, a metadata *registry* (MDR) is a database of metadata that supports the functionality of registration. Registration accomplishes three main goals: identification, provenance, and monitoring quality. Furthermore, an MDR manages the semantics of the metadata, i.e., the relationships (connections) among them.
- *Accessibility* is enabled by "application programming interfaces" (APIs), which allow one to query and retrieve single entries as well as entire archives.
- *Interoperability* implies the use of formal, accessible, shared, and broadly applicable languages for knowledge representation (these are known as formal ontologies and will be discussed in Section "Outlook on ontologies in materials science"), use of vocabularies to annotate data and metadata, and inclusion of references.
- *Reusability* hints at the fact that data in materials science are often exploited only once for a focus-oriented research project, and many data are not even properly stored as they turned out to be irrelevant for the focus. In this sense, many data constitute a "waste" that can be recycled, provided that the data can be found and they are properly annotated.

Establishing one or more metadata schemas that are FAIR-data-principles compliant, and that therefore enable the materials-science community to efficiently share the heterogeneously and decentrally produced data, needs to be a community effort. The workshop "Shared Metadata and Data Formats for Big-Data Driven Materials Science: A NOMAD–FAIR-DI Workshop" was organized and held in Berlin in July 2019 to ignite this effort. In the following sections, we describe the identified challenges and first plans, divided into different aspects that are crucial to be addressed in computational materials science.

In the next Section, we describe the identified challenges and first plans for FAIR metadata schemas for computational materials science, where we also summarize as an example the main ideas behind the metadata schema implemented in the Novel-Materials Discovery (NOMAD) Laboratory for storing and managing millions of data objects produced by means of atomistic calculations (both *ab initio* and molecular mechanics), employing tens of different codes, which cover the overwhelming majority of what is actually used in terms of volume-of-data production in the community. We then follow with more detailed sections discussing the specific challenges related to *interoperability* and *reusability* for ground-state calculations (Section "Metadata for ground-state electronic-structure calculations"), perturbative and excited-state calculations (Section "Metadata

for external-perturbation and excited-state electronic-structure calculations"), potential-energy sampling (molecular-dynamics and more, Section "Metadata for potential-energy sampling"), and generalized workflows (Section "Metadata for Computational Workflows") are addressed in detail in the following sections. Challenges related to the choice of file formats are discussed in Section "File Formats". An outlook on metadata schema(s) for experimental materials science and on the introduction of formal ontologies for materials-science databases constitute Sections "Metadata schemas for experimental materials science" and "Outlook on ontologies in materials science", respectively.

Towards FAIR metadata schemas for computational materials science

The materials-science community has realized long ago that it is necessary to structure data by means of metadata schemas. In this Section, we describe the pioneering and recent examples of such schemas, and how a metadata schema becomes FAIR-data-principles compliant.

To our knowledge, the first systematic effort to build a metadata schema for exchanging data in chemistry and materials science is CIF, an acronym that originally stood for Crystallographic Information File, the data-exchange standard file format introduced in 1991 by Hall, Allen and Brown[5,6]. Later, the CIF acronym was extended to also mean Crystallographic Information Framework[7], a broader system of exchange protocols based on data dictionaries and relational rules expressible in different machine-readable manifestations. These include the Crystallographic Information File itself, but also, for instance, XML (eXtensible Markup Language), a general framework for encoding text documents in a format that is meant to be at the same time human and machine readable. CIF was developed by the International Union of Crystallography (IUCr) working party on Crystallographic Information and was adopted in 1990 as a standard file structure for the archiving and distribution of crystallographic information. It is now well established and is in regular use for reporting crystal structure determinations to Acta Crystallographica and other journals. More recently, CIF has been adapted to different areas of science such as structural biology (mmCIF, the macromolecular CIF[8]) and spectroscopy[9]. The CIF framework includes strict syntax definition in a machine-readable form and dictionary defining (meta) data items. It has been noted that the adoption of the CIF framework in IUCr publications has allowed for a significant reduction of the amount of errors in published crystal structures[10,11].

An early example of an exhaustive metadata schema for chemistry and materials science is the Chemical Markup Language (CML[12–14]), whose first public version was released in 1995. CML is a dictionary, encoded in XML for chemical metadata. CML is accessible (for reading, writing, and validation) via the Java library JUMBO (Java Universal Molecular/Markup Browser for Objects[14]). The general idea of CML is to represent with a common language all kinds of documents that contain chemical data, even though currently the language–as of the latest update in 2012[15]–covers mainly the description of molecules (e.g., IUPAC name, atomic coordinates, bond distances) and of inputs/outputs of computational chemistry codes such as Gaussian03[16] and NWChem[17]. Specifically, in the CML representation of computational chemistry calculations[18], (ideally) all the information on a simulation that is contained in the input and output files is mapped onto a format that is in principle independent of the code itself. Such information is:

- Administrative data like the code version, libraries for the compilation, hardware, user submitting the job;
- Materials-specific (or materials-snapshot-specific) data like computed structure (e.g., atomic species, coordinates), the physical method (e.g., electronic exchange-correlation treatment, relativistic treatment), numerical settings (basis set, integration grids, etc.);
- Computed quantities (energies, forces, sequence of atomic positions in case a structure relaxation or some dynamical propagation of the system is performed, etc…).

The different types of information are hierarchically organized in *modules*, e.g., *environment* (for the code version, hardware, run date, etc.), *initialization* (for the exchange correlation treatment, spin, charge), *molgeom* (for the atomic coordinates and the localized basis set specification), *finalization* (for the energies, forces, etc.). The most recent release of the CML schema contains more than 500 metadata-schema items, i.e., unique entries in the metadata schema. It is worth noticing that CIF is the dictionary of choice for the crystallography domain within CML.

Another long-standing activity is JCAMP-DX (Joint Committee on Atomic and Molecular Physical Data - Data Exchange)[19], a standard file format for exchange of infrared spectra and related chemical and physical information that was established in 1988 and then updated with IUPAC recommendations until 2004. It contains standard dictionaries for infrared spectroscopy, chemical structure, NMR[20], and mass[21] and ion-mobility spectrometry[22]. The European Theoretical Spectroscopy Facility (ETSF) File Format Specifications were proposed in 2007[23–25], in the context of the European Network of Excellence NANOQUANTA, in order to overcome widely known portability issues of input/output file formats across platforms. The Electronic Structure Common Data Format (ESCDF) Specifications[26] is the ongoing continuation of the ETSF project and is part of the CECAM Electronic Structure Library, a community-maintained collection of software libraries and data standards for electronic-structure calculations[27].

The largest databases of computational materials-science data, AFLOW[28], Materials Cloud[29], Materials Project[30], the NOMAD Repository and Archive[31–33], OQMD[34], and TCOD[35] offer application programming interfaces (APIs) that rely on dedicated metadata schemas. Similarly, AiiDA[36–38] and ASE[39], which are schedulers and workflow managers for computational materials-science calculations, adopt their own metadata schema. OpenKIM[40] is a library of interatomic models (force fields) and simulation codes that test the predictions of these models, complemented with the necessary first-principles and experimental reference data. Within OpenKIM, a metadata schema is defined for the annotation of the models and reference data. Some of the metadata in all these schemas are straightforward to map onto each other (e.g., those related to the structure of the

studied system, i.e., atomic coordinates and species, and simulation-cell specification), others can be mapped with some care. The OPTIMADE (Open Databases Integration for Materials Design[41]) consortium has recognized this potential and has recently released the first version of an API that allows users to access a common subset of metadata-schema items, independent of the schema adopted for any specific database/repository that is part of the consortium.

In order to clarify how a metadata schema can explicitly be FAIR-data-principles compliant, we describe as an example the main features of the *NOMAD Metainfo*, onto which the information contained in the input and output files of atomistic codes, both *ab initio* and force-field based, is mapped. The first released version of the *NOMAD Metainfo* is described in ref. [26] and it has powered the NOMAD Archive since the latter went online in 2014, thus predating the formal introduction of the FAIR-data principles[3].

Here, we give a simplified description, graphically aided by Fig. 1, which highlights the hierarchical/modular architecture of the metadata schema. The *elementary mode* in which an atomistic materials-science code is run (encompassed by the black rectangle) yields the computation of some observables (*Output*) for a given *System*, specified in terms of atomic species arranged by their coordinates in a box, and for a given physical model (*Method*), including specification of its numerical implementation. Sequences or collections of such runs are often defined via a *Workflow*. Examples of workflows are:

- Perturbative physical models (e.g., second-order Møller–Plesset, MP2, Green's function based methods such as G_0W_0, random-phase approximation, RPA) evaluated using self-consistent solutions provided by other models (e.g., density-functional theory, DFT, Hartree-Fock method, HF) applied on the same *System*;
- Sampling of some desired thermodynamic ensemble by means of, e.g., molecular dynamics;
- Global- and local-minima structure searches;
- Numerical evaluations of equations of state, phonons, or elastic constants by evaluating energies, forces, and possibly other observables;
- Scans over the compositional space for a given class of materials (high-throughput screening).

The workflows can also be nested, e.g., a scan over materials (different compositions and/or crystal structures) contains a local optimization for each material and extra calculations based on each local optimum structure such as evaluation of phonons, bulk modulus, or elastic constants, etc.

The *NOMAD Metainfo* organizes metadata into sections, which are represented in Fig. 1 by the labeled boxes. The sections are a *type* of metadata, which group other metadata, e.g., other sections or *quantity*-type metadata. The latter are metadata related to scalars, tensors, strings, which represent the physical quantities resulting from calculations or measurements. In a relational-database model, the sections would correspond to tables, where the *data objects* would be the rows, and the quantity-type metadata the columns. In its most simple realization, a metadata schema is a *key-value* dictionary, where the key is a name identifying a given metadata. In *NOMAD Metainfo*, similarly to CML, the key is a complex entity grouping the several attributes. Each item in *NOMAD Metainfo* has *attributes*, starting with its *name*, a string that must be globally unique, well defined, intuitive, and as short as possible. Other attributes are the human-understandable *description*, which clarifies the meaning of the metadata, the *parent section*, i.e., the section the metadata belongs to, and the *type*, whether the metadata is, e.g., a section or a *quantity*. Another possible *type*, the *category* type, will be discussed below. For the quantity-type metadata, other important attributes are *physical units* and *shape*, i.e., the dimensions (scalar, vector of a certain length, a matrix with a certain number of rows and columns, etc.), and *allowed values*, for metadata that admit only a discrete and finite set of values.

All definitions in the *NOMAD Metainfo* have the following attributes:

- A globally unique qualified name;
- Human-readable/interpretable description and expected format (e.g., scalar, string of a given length, array of given size);
- Allowed values;
- Provenance, which is realized in terms of a hierarchical and modular schema, where each *data object* is linked to all the metadata that concur to its definition. Related to provenance, an important aspect of *NOMAD Metainfo* is its *extensibility*. It stems from the recognition that reproducibility is an empirical concept, thus at any time, new, previously unknown or disregarded metadata may be recognized as necessary. The metadata schema must be ready to accommodate such extensions seamlessly.

The representation in Fig. 1 is very simplified for tutorial purposes. For instance, a workflow can be arbitrarily complex. In particular, it may contain a hierarchy of sub-workflows. In the currently released version of the *NOMAD Metainfo*, the elementary-code-run modality is fully supported, i.e., ideally all the information contained in a code run is mapped onto the metadata schema. However, the workflow modality is still under development. An important implication of the hierarchical schema is the mapping of any (complex) workflow onto the schema. That way, all the information obtained by its steps is stored. This is achieved by parsers, which have been written by the NOMAD team for each supported simulation code. One of the outcomes of the parsing is the assignment of a PID to each parsed *data object*, thus allowing for its localization, e.g., via a URI.

The *NOMAD Metainfo* is inspired by the CML, in particular in being hierarchical/modular. Each instance of a metadata-schema is uniquely identified, so that it can be associated with a URI for its convenient accessibility. An instance of a metadata schema can be generated by using a dedicated parser by pairing each parsed value with its corresponding metadata label. As an example, in Listing 1, we show a portion of the YAML file (see section "File Formats") instantiating Metainfo for a specific entry of the NOMAD Archive. This entry can be

searched by typing "entry_id = zvUhEDeW43JQjEHOdvmy8pRu-GEq" in the search bar at https://nomad-lab.eu/prod/v1/gui/search/entries. In Listing 1, key-value pairs are visible as well as the nested-section structuring.

Listing 1. A portion of a YAML file instantiating Metainfo for one entry of the NOMAD Archive.

```
run:
  - program:
      name: VASP
      version: 4.6.35
  - method:
      - dft:
          xc_functional:
            exchange:
              - name: GGA_X_PBE
            correlation:
              - name: GGA_C_PBE
```

The modularity and uniqueness together allow for a straightforward extensibility including customization, i.e., introduction of metadata-schema items that do not need to be shared among all users, but may be used by a smaller subset of users, without conflicts.

In Fig. 1, the solid arrows stand for the relationship *is contained in* between section-type metadata. A few examples of quantity-type metadata are listed in each box/section. Such metadata are also in an *is-contained-in* relationship with the section they are listed in. The dashed arrows symbolize the relationship *has reference in*. In practice, in the example of an *Output* section, the quantity-type metadata contained in such a section are evaluated for a given system described in a *System* section and for a given physical model described in a *Method* section. So, the section *Output* contains a reference to the specific *System* and *Method* sections holding the necessary input information. At the same time, the *Output* section *is contained in* a given *Atomistic-code run* section. These relationships among metadata already build a basic ontology, induced by the way computational data are produced in practice, by means of workflows and code runs. This aspect will be reexamined in Section "Outlook on ontologies in materials science".

We now come to the *category-type* metadata that allow for complementary, arbitrarily complex ontologies to be built by starting from the same metadata. They define a concept, such as "energy" or "energy component", in order to specify that a given quantity-type metadata has a certain meaning, be it physical (such as "energy") or computer-hardware related, or administrative. To the purpose, each section and quantity-type metadata is related to a category-type metadata, by means of an *is-a* kind of relationship. Each category-type metadata can be related to another category-type metadata by means of the same *is-a* relationship, thus building another ontology on the metadata, which can be connected with top-down ontologies such as EMMO[42] (see section "Outlook on ontologies in materials science" for a short description of EMMO).

The current version of *NOMAD Metainfo* includes more than 400 metadata-schema items. More specifically, these are the *common* metadata, i.e., those that are code-independent. Hundreds more metadata are code-specific, i.e., mapping pieces of information in the codes' input/output that are specific to a given code and not transferable to other codes. The *NOMAD Metainfo* can be browsed at https://nomad-lab.eu/prod/v1/gui/analyze/metainfo.

To summarize, the *NOMAD Metainfo* addresses the FAIR-data principles in the following sense:

- *Findability* is enabled by unique names and a human-understandable description;
- *Accessibility* is enabled by the PID assigned to each metadata-schema item, which can be accessed via a RESTful[43] API (i.e., an API supporting the access via web services, through common protocols, such as HTTP), specifically developed for the *NOMAD Metainfo*. Essentially all NOMAD data are open access and users who wish to search and download data do not need to identify themselves. They only need to accept the CC BY license. Uploaders can decide for an embargo. These data are then shared with a selected group of colleagues.
- *Interoperability* is enabled by the extensibility of the schema and the category-type metadata, which can be linked to existing and future ontologies (see Section "Outlook on ontologies in materials science").
- *Reusability/Repurposability/Recyclability* is enabled by the modular/hierarchical structure that allows for accessing calculations at different abstraction scales, from the single observables in a code run to a whole complex workflow (see Section "Metadata for Computational Workflows").

The usefulness and versatility of a metadata schema are demonstrated by the multiple access modalities it allows for. The *NOMAD Metainfo* schema is the basis of the whole NOMAD Laboratory infrastructure, which supports access to all the data in the NOMAD Archive, via the NOMAD API (also an implementation of the OPTIMADE API[41] is supported). This API powers three different access modes of the Archive: the *Browser*[44], which allows searches for single or groups of calculations, the *Encyclopedia*[45], which display the content of the Archive organized by *materials*, and the *Artificial-Intelligence (AI) Toolkit*[46–48], which connects in Jupyter notebooks script-based queries and AI (machine-learning, data-mining) analyses of the filtered data. All the three services are accessible via a web browser running the dedicated GUI offered by NOMAD.

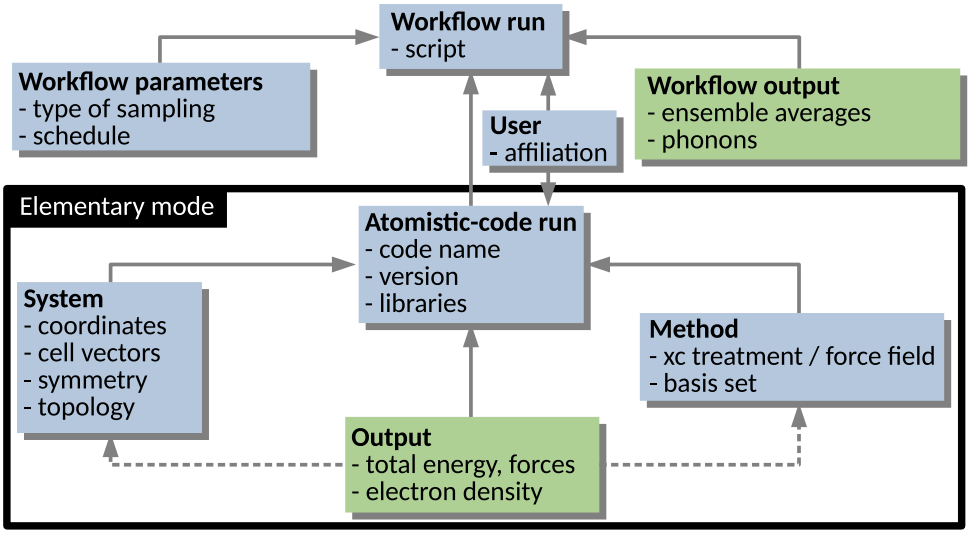

Fig. 1 Simplified schema of the *NOMAD Metainfo*. The rectangles symbolize the section-type metadata, for each section a few examples of therein contained quantity-type or (sub)sections metadata are listed. Sections are always written in bold font. The solid arrows stand for the *is contained in* relationship, while the dashed arrows are for the *has reference* in relationship.

Metadata for ground-state electronic-structure calculations

By ground-state calculations, we mean calculations of the electronic structure–e.g., eigenvalues and eigenfunctions of the single-particle Kohn-Sham equations, the electron density, the total energy and possibly its derivatives (forces, force constants)–for a fixed configuration of nuclei. This refers to a point located on the Born-Oppenheimer potential-energy surface, and is a necessary step in geometry optimization, molecular dynamics, the computation of vibrational (phonon) spectra or elastic constants, and more. Thus, ground-state calculations represent the most common task in computational materials science, and the involved approximations are relatively well established. For this reason, they are already extensively covered by the *NOMAD Metainfo*. Nevertheless, some challenges in defining metadata for such calculations still remain, as discussed below. In particular, density-functional theory (DFT) is the workhorse approach for the great majority of ground-state calculations in materials science. Highly accurate quantum-chemistry models are more computationally expensive than DFT and their use in applications is less widespread. However, they can provide accurate benchmark references for DFT, making high-quality quantum-chemical data essential also for DFT-based studies. Below we analyze the ground-state electronic structure calculations mainly in reference to DFT, but most of the stated principles are also valid for quantum-chemical calculations. A detailed discussion of the latter is deferred to Section "Quantum-chemistry methods".

Approximations to the DFT exchange-correlation functional. Approximations to the DFT exchange-correlation (xc) functionals are identified by a name or acronym (e.g., "PBE"), although sometimes this identification is not unique or complete. As metadata, we suggest to use the identifiers of the Libxc library[49,50], which is the largest bibliography of xc functionals. In order to be both human and computer friendly, the Libxc identifiers consist of a human-readable string that has a unique integer associated with it. Often, the above-noted identification needs some refinement, because xc functionals typically depend on a set of parameters and these may be modified for a given calculation. Obviously, there is a need to standardize the way in which such parameters are referenced. Just like it is possible to use the Libxc identifiers for the functionals themselves, one may also use the Libxc naming scheme for their internal parameters. Obviously, code developers have to ensure that this information is contained in the respective input and/or output files. As Libxc provides version numbers of the xc functionals, it is important that this information is also available.

Basis sets. Complete and unambiguous specification of the basis set is crucial for judging the precision of a calculation. Ground-state calculations should include the full information about the basis sets used, including a DOI that a basis may be referred to. The use of repositories of basis sets, like the Basis Set Exchange repository[51], is therefore strongly recommended.

Basis sets can be coarsely divided into two classes, i.e., atom-position-dependent (atom-centered, bond-centered) and cell-dependent (such as plane waves) ones. Also a combination of both is possible, as, e.g., realized in augmented plane-wave or projector-augmented-wave methods. For the atom-centered basis, the list of centers needs to be provided, and these may even contain positions where no actual atomic nucleus is located. The *NOMAD Metainfo* contains a rather complete set of metadata to describe atom-centered basis sets. A more complete description of cell-dependent basis sets can be found in the ESCDF, which is planned to be merged with the *NOMAD Metainfo*.

Energy reference. In order to enable *interoperability* and *reusability* of energies computed with different electronic-structure methods, it is necessary to define a "general energy zero". An analysis of this problem and some clues on how to tackle it were already discussed by some of us in a previous work[26]. The following is a further attempt to advance and systematize ideas and solutions.

The problem of comparing energies is not restricted to computational materials science and chemistry. In fact, it also arises in experimental chemistry, as for instance, only enthalpy or entropy differences can be measured, but not absolute values. To solve this, chemists have defined a reference state for each element, called the *standard state*, which is defined as the element in its natural form at standard conditions, while the *heat of formation* is used to measure the change from the elements to the compound. In computational materials science and chemistry, we can adopt a similar approach. For each element we need to define a reference system as the zero of the energy scale. To do so, we introduce some definitions:

- A *system* is a defined set of one or more atoms, with a given geometry and, if periodic, a given unit cell. It can be an atom, a molecule, a periodic crystal, etc. If relevant, the charge, the spin-state or magnetic ordering needs to be specified.
- A *reference system* is a well-defined system to which other systems are compared to.
- A *calculated energy* is the energy obtained by a numerical simulation of a system with given input data and parameters, defining the Hamiltonian (i.e., DFT xc-functional approximation) or the many-electron model (e.g., Hartree-Fock, MP2, "coupled-cluster singles, doubles, and perturbative triples", CCSD(T)), the basis set, and the numerical parameters.

Whether the reference system is an atom, an element in its natural form, some molecule or other system, does not matter, as long as it is well defined. Defining the system by atoms requires specifying how the orbitals are occupied, whether the atom is spherical, spin-polarized, etc. For each computational method and numerical settings, the energy per atom of the reference system must be calculated. The *standard energy* is then obtained by subtracting these values (multiplied by the number of constituents) from the calculated total energy. For example, to determine the energy of formation of a molecule like H_2O or a crystal like SiC, we calculate the difference in total energies $E(H_2O) - E(H_2) - \frac{1}{2}E(O_2)$, or $E(SiC) - E(Si) - E(C)$, respectively. Here, H_2 and O_2 are isolated, neutral molecules while Si and C are free, neutral atoms. However, using the energy per atom of Si and C in their crystalline ground-state structure would be an option as well. We propose to tabulate the reference energies for the most common computational methods, so that they can be applied without further computations and preferably automatically by the codes themselves.

Finally, we need to define what is meant by a computational method. The Hamiltonian and DFT functional are clearly part of the definition as is the basis set and the potential shape (including pseudopotentials (PP) and effective core potentials). The specific implementation may also be relevant. Gaussian-based molecular-orbital codes may give the same energy for an identical setup (see Section "Quantum-chemistry methods"), while plane-wave DFT codes may not.

One factor here is the choice of the PP. Irrespective of the used method, the computational settings determine the quality of a calculation. Most decisive here is the basis-set cut-off. For the plane-wave basis, convergence with respect to this parameter is straightforward. In any case, depending on the code, the method and details of the calculation, care needs to be taken to define all the adjustable parameters that significantly affect the energy when defining computational methods.

To tabulate standard energies, as suggested above, every computational method needs to be applied to all reference systems. This requires care in choosing the reference systems to ensure that an as-wide-as-possible range of codes and methods are actually suited for these calculations. It may be that some codes cannot constrain the occupancies of atoms, or keep them spherical, which would be a problem if spherical atoms were chosen as the reference. Clearly, periodic crystals such as silicon are not suitable for molecular codes. It is possible, however, that some other codes could help bridging this gap. For example, FHI-aims[52] is not only capable of simulating crystalline system, but can also handle atoms and molecules and it can employ Gaussian-type orbitals (GTO) basis sets. Thus, FHI-aims is able to reproduce energy differences between atoms/molecules and crystals. In this way, it can support codes such as Gaussian[16] or GAMESS[53].

Metadata for external-perturbation and excited-state electronic-structure calculations

A direct link from the DFT ground state (GS) to excitations is provided by time-dependent DFT (TDDFT). Alternatively, charged and neutral electronic excitations are described by means of Green-function approaches from many-body perturbation theory (MBPT). This route is predominantly (but not exclusively) used for the solid state, while TDDFT and quantum-chemistry approaches are typically preferred for finite systems. For strongly correlated materials, in turn, dynamical mean-field theory (DMFT) is often the methodology of choice, potentially combined with DFT and Green-function methods. Lattice excitations, if not directly treated by DFT molecular dynamics, are often handled by density-functional perturbation theory (DFTP); for their interaction with light, also Green-function techniques are used. DFPT not only allows for the description of vibrational properties, but also for treating macroscopic electric fields, applied macroscopic strains, or combinations of these. The type of perturbation is intimately related to the physical properties of interest, e.g., harmonic and anharmonic phonons, effective charges, Raman tensors, dielectric constants, hyper-polarizabilities, and many others.

Characterizing the corresponding research data is a very complex and complicated task, for various reasons. First, such calculations rely on an underlying ground-state calculation, and thus carry along all uncertainties

from it. Second, the methodology for excited states is scientifically and technically more involved by including many-body effects that govern diverse interactions. The methods thus rely on various, often not fully characterized approximations.

Diagrammatic techniques and TDDFT. The most common application of *GW* is to compute quasi-particle energies, i.e., energies that describe the removal or addition of a single electron. For this, the many-body electron-electron interaction is described by a two-particle operator, called the electronic self-energy. To compute this object, on the technical side we may need an additional (auxiliary) basis set, not the same as the one used in the ground-state calculation, coming with additional parameters. Likewise, there are various ways for doing the analytical continuation of the Green's function, as there are various ways for carrying out the required frequency integration, possibly employing a plasmon-pole model as an approximation. And there are also different ways of how to evaluate the screened Coulomb potential *W*. Most important is the flavor of *GW*, i.e., whether it is done in a single-shot manner, called G_0W_0, or in a self-consistent way. If the latter, what kind of self-consistency (scf) is used —any type of partial scf, quasi-particle scf, or any other type which would remedy any starting-point dependence, i.e., the dependence of the results on the xc functional of the initial DFT (or Hartree-Fock or alike) used in the GS.

While *GW* is the method of choice for quasi-particle energies (and potentially also life times) within the realm of MBPT, we need to solve the Bethe-Salpeter equation (BSE) to tackle electron-hole interactions. This approach should typically be applied on top of a *GW* calculation, but often the quasi-particle states are approximated by DFT results adjusted by a scissors operator to widen the band gap in a similar way to the latter. In all cases, BSE carries along all subtleties from the underlying steps. In addition, it comes with its own issues, like the way of screening the Coulomb interaction (electron-hole this time), the representation of non-local operators, and alike.

DMFT, as a rather young and quickly developing field, naturally experiences a plenitude of "experimental" implementations, differing in many aspects, with one of the major obstacles being the quite vast amount of combinations of software. Some of the approaches are computationally light, allowing for the construction of model Hamiltonians based on DFT calculations; others are computationally too demanding and can be applied only to simple systems with a few orbitals; most of the methods rely on Green's functions and self-energies. Diagrammatic extensions beyond standard DMFT methods employ various kinds of vertex functions. Other issues concern the definition of how to handle the Coulomb interactions, where the parameters can either be chosen empirically or can be calculated by first principles.

Specific issues of TDDFT concern, in a first place, the distinction between the linear-response regime and the time -propagation of the electronic states in presence of a time-dependent potential. For the former, the xc kernel plays the same role as the xc functional of the GS, raising (besides numerical precision) questions related to accuracy. For the latter, there are various ways and flavors for how to implement the time-evolution operator. Moreover, one can write this operator as a simple exponential or use more elaborate expressions, like the Magnus expansion or the enforced time reversal symmetry. Regarding the exponential, one can employ a Crank-Nicolson expansion, expand in a Taylor series or employ Houston states. Obviously, each of them comes with approximations and additionally, numerical issues.

In summary, all the variety captured by the different methods together with the related multitude of computational parameters, needs to be carefully reflected by the metadata schema. This is not only imperative for ensuring reproducible results but also for evaluating the accuracy of methods and commonly used approximations. Besides, further subtleties related to algorithms in the actual implementations in different codes requires the code developers to embark on this challenge.

Density-functional perturbation theory. Density-functional perturbation theory is used to obtain physical properties that are related to the (density-)response of the system to external perturbations, like the displacement potential according to lattice vibrations. Also in this case, the calculation relies on a preliminary GS run, inheriting all issues therefrom. After having chosen the type of perturbation, which requires method-dependent definitions and inputs, one needs to choose the order of perturbation: The linear response approach, that is implemented in many codes (e.g., VASP[54], octopus[55], CASTEP[56], FHI-aims[57], Quantum Espresso[58], ABINIT[59]), allows for the determination of second-order derivatives of the total energy. Among these codes, some of them also allow for the calculation of third-order derivatives, like anharmonic vibrational effects. The variation of the Kohn-Sham orbitals can be obtained from the Sternheimer equation, where different methods are used for deriving its solution (iterative methods, direct linearization, integral formulation).

Quantum-chemistry methods. Quantum chemistry offers several methodological hierarchies for calculating quantities related to excited states, such as excitation energies, transition moments, ionization potentials, etc. As high-quality methods are computationally intensive, without additional approximations such methods can be applied to relatively small molecular systems only.

Among the standard quantum chemical approaches that can be routinely applied to study excited states of small to medium-sized molecules one can distinguish two large groups, i.e., single-reference and multi-reference methods. The single-reference coupled-cluster (CC) hierarchy for excited states can be formulated in terms of the so-called equation-of-motion approach or time-dependent linear response.

Generally, for well-behaving closed-shell molecules, the single-reference quantum-chemical methods can be used as a black box. The formalisms of the MP *n* and CC models are uniquely defined and well documented. The GTO basis sets from the standard basis set families (Pople, Dunning, etc.) are also uniquely defined by the acronym. In practical implementations of these methods, of course various thresholds are usually introduced

for prescreening, convergence, etc., but the default values for these thresholds are routinely set very conservatively to guarantee a sub-microhartree precision of the final total energies. Problems might, however, arise due to the iterative character of most of the mentioned techniques, as convergence to a certain state (both in the ground-state and/or excited-state parts of the calculation) depends on starting guess, preconditioner, possible level shifts, type of convergence accelerator, etc. Unfortunately, the parameters that control the convergence are often not sufficiently well documented and might not be found in the output. Such problems mainly occur in open-shell cases (note that in the Delta methods at least one of the calculations has to involve an open-shell system). Sometimes a cross-check between several codes becomes essential to detect convergence faults.

When it comes to larger systems and approximate CC models are utilized, the importance of the involved tolerances and underlying protocols substantially increases. The approximations can include, for example, the density-fitting technique, local approximation, Laplace transform, and others. Important parameters here are the auxiliary basis set, the fitting metric, the type of fitting (local or non-local), and if local, how the fit domains are determined, etc. The result of the calculations that use local correlation techniques are influenced by the choice of the virtual space and the corresponding truncation protocols and tolerances, the pair hierarchies and the corresponding approximations for the CC terms, etc. For Laplace-transform-based methods, the details of the numerical quadrature matter. Unfortunately, these subtleties are very specific and technical and even if given in the output, can hardly be properly understood and analyzed by non-specialists who are not involved in the development of the related methods. Therefore, the protocols behind the approximations are usually appropriately automatized, and the defaults are chosen such that for certain (benchmarking) sets of systems the deviations in the energy are substantially smaller than the expected error of the method itself (e.g., 0.01 eV for the excitation energy). However, for these methods, additional benchmarks and cross-checks between different programs and approaches would be very important.

Multi-reference methods come with quite a number of different flavors, where the most widely used ones are complete active-space self-consistent Field (CASSCF), complete active-space second-order perturbation theory (CASPT2), and multireference configuration interaction (MRCI). For difficult cases (e.g., strongly correlated systems), these methods might remain the only option to obtain qualitatively and quantitatively correct result. Unfortunately, compared to the single reference methods, they are computationally expensive and much less of a black box. First of all, for each calculation one has to specify the active space or active spaces. The results may depend dramatically on this choice. Furthermore, the underlying theory is not always uniquely defined by the used acronym. For example, different formulations of CASPT2, MRCI, or other theories are not mutually equivalent depending on whether and how much internal contraction is used and additional approximations that neglect certain terms (e.g., many-electron density matrices) can be implicitly invoked. Besides, certain deficiencies of these methods, such as for example lack of size consistency in MRCI or intruder states in CASPT2, are often corrected by additional (sometimes empirical) schemes, which again are not always fully specified. All this makes the interpretation of deviations in results and cross-checks of these methods less conclusive.

To summarize, quantum-chemical methods offer an excellent toolbox for accurate *ab initio* calculations for molecules (especially so for small and medium sized ones). However, severe issues concerning reproducibility and replicability remain, in particular for extended and/or open-shell systems. This calls for a more detailed specification of the implemented techniques by the developers, for example, a better design of the outputs, and a thorough analysis and documentation of the employed methods and parameters by the users. A possible strategy addressing these issues would be two-fold. a) Promoting the compliance of the developed software with the FAIR principles for software[60,61], which comprise the recommendation to publish the software in a repository with version control, have a well-defined license, register the code in a community registry, assign to each version a PID, and enable its proper citation[62,63]. Reproducibility can be enhanced by publishing software code under the Free/Libre Open-Source Software (F/LOSS)[64,65] license and by documenting the computation environment (hardware, operating system version, computational framework and libraries that were used, if any) b) Creation of well-defined benchmark datasets. Interoperability among different implementations of (in the intention) the same theoretical model can be assessed by the quantitative comparison over different codes (including different versions thereof) of a set of properties on an agreed-upon set of materials. Such datasets would obviously need to be stored in a FAIR-data-compliant fashion. A large community-based effort in this direction is being carried on in the DFT community[66], while in the many-body-theory community, implementation of this idea is just at its beginning[67]).

Metadata for potential-energy sampling

Molecular dynamics (MD) simulations model the time evolution of a system. They employ either *ab initio* calculated forces and energies (aiMD) or molecular mechanics (MM) i.e., forces and energies are defined through empirical atomistic and coarse-grained potentials. The FAIR storing and sharing of their inputs and outputs comes with a number of specific challenges in comparison to single-point electronic-structure calculations.

Conceptually, aiMD and MM are similar, as a sequence of system configurations is evolved at discrete time steps. Positions, velocities, and forces at a given time step are used to evaluate positions and velocities, and hence forces in the new configuration, and so on. In practice, MM simulations are orders of magnitude faster than aiMD, enabling much longer time scales and/or much larger system sizes. Even though the trend towards massive parallelization will enable aiMD in the near future system to handle sizes comparable to today's standards for MM simulations, the latter will probably always enable larger systems. However, with machine-learned potentials and active learning techniques for their training, aiMD and MM may grow together in the future.

In this Section, we focus on challenges more specific to MM simulations, having in mind large length scales, long time scales, and complex phase-space-exploration algorithms and workflows. They can be summarized as follows:

(i). In many cases, the investigated systems feature thousands of atoms with complex short- and long-range order and disorder, e.g., describing microstructural evolution such as crack propagation. This requires large, complex simulation cells with a range of chemical species to be correctly described and categorized.
(ii). Force-fields exist in a wide variety of flavors that require proper classification. On top of that, they allow for granular fine-tuning of the interactions, even for individual atoms. Faithfully representing complex force fields thus requires to also capture the chemical-bonding topology that is often needed to define the actual interactions.
(iii). The large length and long time scales presently come together with a multitude of simulation protocols, which use specific boundary conditions, thermostats, constraints, integrators, etc. The various approaches enable the computations of additional observables to be computed as statistical averages or correlations. Representing these properties implies the need to efficiently store and access large volumes of data, e.g., trajectories, including positions, and possibly also velocities and forces, for each atom at each time step.

For the purpose of illustration, we start by identifying some typical use cases, then describe what is currently implemented in the NOMAD infrastructure and what is missing. The examples we adopt fall into two classes: (i) high throughput systems that are individually *simple* (1000–10000 particles) where the value of sharing comes from the ability to run analysis across many variants of, e.g., chemical composition or force field; (ii) sporadic simulations of very large systems or very long time scales which cannot readily be repeated by other researchers and thus are individually valuable to share. Examples of the first class, could be MD simulations in the *NVT* ensemble for liquid butane or bulk silicon, using well-defined standard force fields (e.g., CHARMM or Stillinger-Weber). Quantities of interest are typically computed during MD simulations (e.g., liquid densities). For flexibility, full trajectory files should also be stored but some important observables might be worth precomputing (e.g., radial distribution functions). The second class could include multi-billion atom MD simulations of dislocation formation[68] or solidification[69,70] or very long time-scale simulations of protein folding[71]. For more complex use cases, the current infrastructure as discussed in Section "Towards FAIR metadata schemas for computational materials science" is not yet sufficient. The challenges to be addressed are the need for support for (i) complex, heterogeneous, possibly multi-resolution systems; (ii) custom force fields; (iii) advanced sampling; (iv) classes of sampling besides MD (e.g., Monte Carlo, global structure prediction/search); (iv) larger simulations (i.e., need for sparsification of the stored data with minimal loss of information)

Complex systems include heterogeneous systems, e.g., adsorbate and surfaces, interfaces, solute (macro)molecules in solvent fluids, and multi-resolution systems, i.e., systems that are described at different granularity. The representation of complex systems requires a hierarchy of structural components, from atoms, through moieties, molecules, and larger (super)structures. Annotating such complexity will require human intervention as well as algorithms for automatically recognizing the structural elements (see, e.g., ref. [72]).

Annotation of *force fields* into publicly accessible databases has been pioneered by OpenKIM[40] in materials science and MoSDeF[73] for soft matter. However, many simulations are performed with customized force fields. The field is already being augmented and will likely be further supported by machine-learning (ML) force fields. So far, the great majority of ML force fields are used only in the publication where they are defined. The *reusability*-oriented annotation of force fields, including ML ones, require also establishing a criterion for comparing them. Comparisons can be carried out by means of standardized benchmark datasets, with a well-defined set of properties. Differences among predicted properties can establish a metric for the similarity of the force fields.

Advanced-sampling techniques (e.g, metadynamics[74], umbrella sampling[75], replica exchange[76], transition-path sampling[77], and forward-flux[78] sampling) are typically supported by libraries such as PLUMED[79] and OpenPathSampling[80]. These libraries are used as plugins to codes where classical-force-field-based (e.g. GROMACS[81], DL_POLY[82], LAMMPS[83]) or *ab initio* (e.g., CP2K[84] and Quantum Espresso[58]) MD, or both (e.g., i-Pi[85]), are performed. The input and output of these plugins will serve as the basis for the metadata related to these sampling techniques. In this regard, it would also be interesting to connect materials-science databases, such as the NOMAD Repository and Archive[31] or Materials Cloud Archive[29] to the PLUMED-NEST[86], the public repository of the PLUMED consortium[87], for example by allowing for automatic uploading of PLUMED input files to the PLUMED-NEST when uploading to the data repositories.

For *long time-* and *large length-scale* simulations, several questions arise: How should we deal with these simulations, where the extensive amount of data produced by MD simulations becomes overwhelmingly large to systematically store and share? Can we afford to store and share all of it? If the storage is limited or data retrieval is unpractically slow, how can we identify the significant and crucial part of the simulation to store it in a reduced form? Keeping the whole data locally and sharing the metadata with only the important parts of the simulations would be a viable alternative, assuming the different servers have enough redundancy. Standard analysis techniques such as similarity analysis and monitoring dynamics can also be used to identify the changes in structure and dynamics to store only the significant frames or specific regions in MD simulations (e.g., some QM/MM models uses large MM buffer-atom regions that may not be stored entirely). Furthermore, on the one hand the cost/benefit of storing versus running a new simulation must be weighed. On the other hand, researchers may soon face increased requirements from funding agencies to store their data for a number of years, in which case the present endeavour offers a convenient implementation. We note ongoing algorithmic developments on compression algorithms for trajectories, see, e.g., ref. [88].

Metadata for Computational Workflows

A computational workflow represents the coordinated execution of *repeatable* (computational) steps while accounting for *dependencies* and *concurrency* of tasks. In other words, a workflow can be thought as a script, a wrapper code that manages the scheduling of other codes, by controlling what should run in parallel, what sequentially and/or iteratively. This definition can be extended to workflows in experimental materials science or hybrid computational-experimental investigations, but, consistently with the previous sections, we limit the discussion to computational aspects only.

Once shared, workflows become useful building blocks that can be combined or modified for developing new ones. Furthermore, FAIR data can be reused as part of workflows completely unrelated to the workflows with which they were generated. An obvious example is artificial-intelligence-based data analytics, which can entail complex workflows involving data originally created for different purposes. During the last decade, the interest in workflow development has grown considerably in the scientific community[89] and various multi-purpose engines for managing calculation workflows, have been developed, including AFLOW[28,90,91], AiiDA[36,92], ASE[39], and Fireworks[93]. Using these infrastructures, a number of workflows have been used for scientific purposes, like convergence studies[94], equations of state (e.g., AFLOW Automatic Gibbs Library[95] and the AiiDA common workflows ACWF[96]), phonons[97–100,101], elastic properties (e.g., the elastic-properties library for Inorganic Crystalline Compounds of the Materials Project[102], AFLOW Automatic Elasticity Library, AEL[103], ElaStic[104]), anharmonic properties (e.g., the Anharmonic Phonon Library, APL[105], AFLOW Automatic Anharmonic Phonon Library, AAPL[106]), high-throughput in the compositional space (e.g., AFLOW Partial Occupation, POCC[107]), charge transport (e.g., organic semiconductors[108,109]), of covalent organic frameworks (COFs) for gas storage applications[110], of spin-dynamics simulations[111], high-throughput automated extraction of tight-binding Hamiltonians via Wannier functions[112], and high-throughput on-surface chemistry[113].

There are two types of metadata associated to workflows. Thinking of a workflow as a code to be run, the first type of metadata characterizes the code itself. The second type is the annotation of a run of a workflow, i.e., its inputs and outputs. This type of metadata has been already described in Section "Towards FAIR metadata schemas for computational materials science", together with a schematic list of possible workflow classes. It is important to realize that the inputs and outputs of the elementary-mode runs of the atomistic codes that are invoked in a workflow run are complemented by the inputs and outputs of the overarching workflows. A simple example: In an equation-of-state type of workflow, the energy and volume per unit cell of each single configuration that is part of the workflow is the output of the elementary run of the code, while the energy-vs-volume equation of state, e.g., fit to the Birch-Murnagham model, is an output of the workflow.

File Formats

On an abstract level, a metadata schema is independent from its representation in computer memory, on a hard drive, or on just a piece of paper. But on a practical level, all data and metadata need to be managed, i.e., stored, indexed, accessed, shared, deleted, archived, etc. File formats used in the community address different requirement and intended use cases. Some file formats privilege human readability (e.g., XML, JSON, YAML) but are not very storage efficient, others are binary and overall optimized for efficient searches, but require interpreters to be understood by a person (e.g., HDF5[114]). There are a few use-cases and data properties in the domain of computational materials science that are worth mentioning. First, such data are very heterogeneous and contain many simple properties (e.g., the name of a used code, or a list of considered atoms) that are mixed with properties in the form of large vectors, matrices, or tensors (e.g., the density of states or wave functions). The number of different properties requires hierarchical organization (e.g., with XML, JSON, YAML, or HDF5). It is desirable that many properties are easily human readable (e.g., to quickly verify the sanity of a piece of data), on the other hand large matrices should be stored as efficiently as possible for archiving, retrieving, and searching purposes. Second, there are use cases where random (non-sequential) access of individual properties is desirable (e.g., return all band structures from a set of DFT calculations). Third, computational-material-science (meta)data need to be archived (efficient storage, prevention of corruption, backups, etc.) on one side, but they also need to be shared via APIs, e.g., for search queries. This requires to transform (meta)data from one representation in one file format (e.g., BagIt and HDF5) to another representation in a different format (e.g., JSON or XML).

These use-cases and data properties lead to the following conclusions: Even on a technical level, (meta) data need to be handled independently of the file format. Pieces of information have to be managed in different formats, and we need to be able to transform from one representation into another. If many different resources (files, databases, etc.) are used to store (meta)data from a logically conjoined dataset, references to these resources qualify to become an important piece of metadata itself. We propose to use an abstract interface (e.g., implemented as a Python library) based on an abstract schema. This interface allows to manage (meta) data independent of the actual representation used underneath. Various implementations of such an abstract interface can then realize storage in various file formats and access to databases.

Metadata schemas for experimental materials science

In contrast to computational materials science, in experimental materials science the atomic structure and composition is only approximately known. Several techniques are used to collect data that may be more or less directly interpreted in terms of the atomic and/or electronic structure of the material. In cases where the structure of the material is already known, careful characterization of properties helps to establish valuable relationships between structure and properties which, in turn, may help to refine theoretical models of these structure-properties links. The inherent uncertainty in every measurement process causes the precision with which data can be reproduced to be lower, in most cases, than in theoretical/computational materials science. These uncertainties are present even in a well-characterized experimental setup, i.e., when a comprehensive

set of metadata is used. In many cases it is not even the focus of an experiment to produce the most perfectly characterized data, but to invest just enough effort to address the specific question that drives the experiment.

The information available about the material whose properties are to be measured is also much less complete than in the computational world, where often the position of every atom is known. However, while physical measurements may be limited in their precision, the accuracy with which a physically observable quantity is obtained is by definition of being physically observable much higher than in computational materials science, where the accuracy of the obtained physical quantity may depend strongly on the validity of approximations being applied.

The uncertainty in retrieving structure-property relationships in computational materials science, which depends on the suitability of the applied theoretical model and its computational implementation, translates in the realm of experiments to an uncertainty in the atomic structure of the object that is being characterized and generally also some uncertainty in the measurement process itself. The metadata necessary to reproduce a given experimental data set must thus include detailed information about the material and its history together with all the parameters which are required to describe the state of the instrument used for the characterization. In most cases, both classes of metadata, i.e., those describing the material and those describing the instrument are going to be incomplete. While, for example, the full history of temperature, air pressure, humidity, and other relevant environmental parameters are not commonly tracked for the complete lifetime of a material (counter-examples exist, e.g., in pharmaceutical research), also information about the state of the instrument is not generally as comprehensive as it should ideally be (e.g., parameters are not recorded, or are not properly controlled, such as hysteresis effects in devices involving magnetic fields, or many mechanical setups).

To overcome part of the uncertainty in the data, one needs to collect as many metadata about the material and its history, as possible, including those that one has no immediate use for at the moment, but might potentially need in the future. Since most of the research equipment being used for characterization tasks is commercial instrumentation, collecting this metadata in an (ideally) fully automated fashion requires the manufacturer's support. In many cases the formats in which scientific data are provided by these instruments is proprietary. Even if all the data to describe the instrument's condition of operation are stored, large parts of them may get lost when using the vendor's software to export the data to other formats; mostly because the "standard format" does not foresee storing vendor- and instrument-specific metadata. It is however worth mentioning here that the CIF dictionaries (see section "Towards FAIR metadata schemas for computational materials science") already contain (meta)data names to describe instrumentation, sample history, and standard uncertainties in both measured and computed values. As a useful addition, the CIF framework provides tools for implementing quality criteria, which can be used for evaluating the trustworthiness of data objects. In this respect, the community has been developing with CIF a powerful tool onto which a FAIR representation of at least structural data can be built.

At large research infrastructures like synchrotrons and neutron-scattering facilities, where a significant fraction of instruments is custom built, and data are often shared with external partners, standards for file formats and metadata structures are being agreed upon, a prominent one being the NeXus standard. NeXus[115] defines hierarchies and rules on how metadata should be described and allows compliant storage using HDF5. Experimental research communities can profit from these activities and provide *NeXus-format application definitions* which describe necessary metadata that should be stored in a dataset, along with definitions for some optional metadata. This common file format for scientific data is slowly beginning to spread to other communities. Having a standard file format for different types of scientific data seems to be an important step forward towards FAIR data management, since it severely reduces the threshold to share data across communities. Note that NeXus provides a glossary and connected ontology which helps in machine interpretability and so in *reusability*.

While standard file formats are of very high value in making data *findable* and *accessible*, due to common use of keywords to describe a given parameter, they also make them more *interoperable*, since the barrier for reading the data is lowered. However, making experimental data truly reproducible requires in many cases more metadata to be collected. Only if the uncertainty with which data can be reproduced is well understood, they may also be fully *reusable*. As discussed in the previous paragraph, part of these metadata must be provided by manufacturers of commercially available components of the experimental setup. Often this just requires more exhaustive data export functions and/or proper, i.e. versioned descriptions, for all of the instrument-state-describing metadata which are being collected during the experiment. Additionally, it may be necessary to equip home-built lab equipment with additional sensors and functionalities for logging their signals.

Even with added sensors and automated logging of all accessible metadata, in many cases, it is also necessary to compile and complete the record of metadata describing the current and past states of the sample that is being characterized by manually adding information and/or combining data from different sources. Tools for doing this in a machine-readable fashion are Electronic Lab Notebooks (ELNs) and/or Laboratory Information Management Systems (LIMS). Many such systems are already available[116–122], including open-source solutions that combine features of both ELN and LIMS into one software. Server-client solutions that do not require a specific client, but may be accessed through any web browser, have the advantage that information may be accessed and edited from any electronic device capable of interacting with the server. Such ease of access, combined with the establishment of rules and practices of holistic metadata recording about sample conditions and experimental workflows will also help to increase the reproducibility and with that the *reusability* of experimental data. The easier the use of such a system is, and the more apparent it makes the benefits of the availability of FAIR experimental data, the faster it will be adopted by the scientific community.

Outlook on ontologies in materials science

In data science, an ontology is a *formal representation of the knowledge of a community about a domain of interest, for a purpose*. As ontologies are currently less common in basic materials science than in other fields of science, let us explain these terms:

- *Formal representation* means that: (1) the ontology is a *representation*, hence it is a simplification, or a model, of the target domain, and (2) the attribute *formal* communicates that the ontological terms and relationships between them must have a deterministic and unambiguous meaning. Furthermore, *formal representation* implies that the mechanism to specify the ontology must have a degree of logical processing capability, e.g., inference and reasoning should be possible. Crucially, the attribute *formal* refers to the fact that an ontology should be machine readable.
- *Knowledge* is the accumulated set of facts, pieces of information, and skills of the experts of the domain of interest that are represented in the ontology.
- The *community* influences the ontology in two aspects; (1) it implies an overall agreement between a group of experts/users of the knowledge as represented in the ontology and (2) it indicates that the ontology is not meant to convince a whole population nor wants to be universal. However, if it fulfills the requirements of bigger communities, the ontology will be adopted by broader audiences and will find its way towards standardization.
- The *domain of interest* is the common ground for the community, e.g., a scientific discipline, a subordinate of discipline, or a market section. It is often used as a boundary to limit the scope of the ontology. It is a proper tool to detect overlapping concepts, modularizing ontologies, and identifying extension and integration points.
- The *purpose* conveys the goals of the ontology designers so that the ontology is applicable to a set of situations. In many ontology design efforts, the purpose is formulated by a collection of so-called *competency questions*. These questions and the answers provided to them identify the intent and viewpoint of the designers and set the potential applications of the ontology.

In practice, ontologies are often mapped onto, and visualized by means of, directed acyclic graphs, where an edge is one of a well-defined set of relationships (e.g., *is a*, *has property*) and each node is a *class*, i.e., a concept which is specific to the domain of interest. Each node-edge-node *triple* is interpreted as a subject-predicate-object expression. For instance, in an ontology for catalysis, one could find the triples: "catalytic material–has property–selectivity", and "selectivity–refers to–reaction product". Ontologies address the *interoperability* requirement of FAIR data. By means of a machine-readable formal structure, which can be connected to an existing or *ex novo* derived metadata schema of a database, ontologies allow queries over various databases, even from different fields.

The literature already contains several ontologies created for representing (aspects of) materials science. The most ambitious project is probably EMMO[42], which stands for both European Materials Modelling Ontology, developed within the European Materials Modelling Council (EMMC), and Elemental Multiperspective Material Ontology. EMMO is designed to provide a formal way to describe the fundamental concepts of physics, chemistry, and materials science, to provide an all-purposes common ground for describing materials, models, and data that can be adapted by all sub-domains of condensed-matter physics and chemistry. The development of EMMO includes also a handful of *domain ontologies* that assume EMMO as top-level ontology[123]. These domain ontologies span subjects such as "atomistic and electronic modeling", "crystallography", "mechanical testing", and more. So far, however, EMMO and its domain ontologies have not been connected to existing databases.

Other domain-specific ontologies, not related to EMMO, have been developed. For instance, the Materials Ontology[124] was developed for the exchange of data among databases for thermal properties, the MatOnto ontology[125] addresses oxygen ion conducting materials in the fuel cell domain, the NanoParticle Ontology[126] maps properties of nanoparticles with the purpose of designing new nanoparticles with given properties, while the eNanoMapper ontology[127] focuses on assessing risks related to the use of nanomaterials from the engineering point of view.

An application-oriented ontology is Materials Design Ontology (MDO)[128], developed under the guidance of the schemas from OPTIMADE[41], and therefore aimed at dealing with data from the various materials-data repositories (AFLOW, Materials Project, etc.) on a common ground. In practice, MDO connects calculated structures with the calculated properties and the physical model adopted to calculate structures and properties. Furthermore, the provenance for each calculation, is also represented in MDO. It has recently been extended using text mining on thousands of journal articles[129].

The hierarchical structure of *NOMAD Metainfo* already includes ontological aspects. More specifically, it represents atomistic calculations, as performed by all the parsed simulation codes. *NOMAD Metainfo* contains already five types of relations between the metadata: (a) is subclass of, (b) is part of, (c) has reference, (d) has dimension and (e) has category. The latter relation, *has category* is introduced to describe conceptually physical quantities (e.g., "energy", "velocity", etc.). Recently[130], this basic *NOMAD Metainfo* ontology has been expanded to include a representation of operations among arrays (in an ontology, any mathematical concept needs to be represented in order to properly operate with the physical quantities in complex queries). This extension allowed for the introduction of the notion of "similarity" relationship that has been applied as a proof of concept to the calculated electronic density of states, as stored in the NOMAD Archive, in order to identify materials with similar electronic structures[131,132].

Achievements and challenges of ontologies for materials science were discussed at the first "Workshop on Ontologies for Materials-Databases Interoperability (OMDI2021)", held in Linköping and virtually on October 2021. The workshop was organized by the OPTIMADE consortium[41] and funded by Psi-k[133]. The main outcomes of the workshop were: a) the strengthening of the idea that the development of useful ontologies need a community effort; b) they need to build from the data, i.e., their development needs to be driven by existing data and the aim of connecting data from different sources; c) tools for text mining need to be developed[129,134], in order to map into ontologies the enormous wealth buried in decades of scientific literature. Another important outcome of the workshop was the utterance of an insightful warning: "is the field proposing solutions (i.e, the existing ontologies) still in search of a problem?". In other words, the community realizes that it needs specific questions to be addressed (the competence questions) in order to shape the ontologies and then propose demonstrative applications of such ontologies to answer the agreed upon questions.

Discussion and Outlook

Defining–as completely as possible–a pool of metadata for all the methods and computed quantities described above, is crucial for processing, storing, and providing FAIR materials-science data. A key challenge is the mapping into a metadata schema of the full set of input parameters, including those hidden into the specific codes, and all the available output. This practice will facilitate reproducibility, benchmarking, and peer-review processes.

In particular, we emphasize the importance of developing a hierarchical and modular metadata schema in order to represent the complexity of materials science data and allow for access, reproduction, and repurposing of data, from single-structure calculations to complex workflows. Furthermore, the modularity of the schema enables its extensibility, which is vital for the long-term maintenance of the metadata infrastructure.

As an example, we presented the current status of the NOMAD metadata schema, which was designed to comply with the FAIR principles. By means of existing parsers that map a growing set of atomistic-simulation code packages into the hierarchical, modular NOMAD metadata schema, the NOMAD infrastructure already provides the community with a FAIR storage of materials science data. The challenges of fully covering the ground-state electronic calculations, and extending the schema to excited states, dynamical simulations, and complex workflows were examined in detail. By means of a community effort, all aspects of the different subfields, and all the practical details of each specific implementation can be mapped on the NOMAD metadata schema. Finally, we discussed the challenges of the *FAIRification* of experimental materials-science metadata and the creation of ontologies for materials science. Ontologies will unlock the *interoperability* of the FAIR data by enabling the access and reuse of data across materials-science areas, but also outside materials science.

As a perspective, probably the biggest benefit of meeting the interoperability challenge will be to allow for routine comparisons between computational evaluations and experimental observations. In fact, it is not trivial to associate a given computed quantity, derived through a given theoretical modelling, to an experimentally measured quantity. This association requires the judgment of a domain expert and a full characterization of both compared quantities. This is where a formalized ontology, applied to FAIR data in materials science, could automatize the process.

Received: 30 May 2022; Accepted: 23 August 2023;
Published online: 14 September 2023

References

1. Rickman, J. M., Lookman, T. & Kalinin, S. V. Materials informatics: From the atomic-level to the continuum. *Acta Mater.* **168**, 473–510 (2019).
2. Hey, T., Tansley, S. & Tolle, K. *The Fourth Paradigm: Data-Intensive Scientific Discovery* (Microsoft Research, 2009).
3. Wilkinson, M. D. *et al*. The FAIR guiding principles for scientific data management and stewardship. *Sci. Data* **3**, 1–9 (2016).
4. Grassi, P., Lefkovitz, N., Nadeau, E., Galluzzo, R. & Dinh, A. Attribute metadata: A proposed schema for evaluating federated attributes. Tech. Rep., National Institute of Standards and Technology. https://doi.org/10.6028/NIST.IR.8112 (2018).
5. Hall, S. R., Allen, F. H. & Brown, I. D. The crystallographic information file (CIF): a new standard archive file for crystallography. *Acta Crystallographica Section A: Foundations of Crystallography* **47**, 655–685 (1991).
6. Bernstein, H. J. *et al*. Specification of the crystallographic information file format, version 2.0. *Journal of Applied Crystallography* **49**, 277–284 (2016).
7. Hall, S. R. *et al*. Formal specification of the crystallographic information file. version 1.1 specification. In Hall, S. & McMahon, B. (eds.) *International Tables for Crystallography, Vol. G, Definition and Exchange of Crystallographic Data*, 25–36 (Springer, Dordrecht, 2005).
8. Westbrook, J., Yang, H., Feng, Z. & Berman, H. The use of mmCIF architecture for PDB data management. *International Tables for Crystallography. Dordrecht, The Netherlands: Springer* 539–543 (2005).
9. El Mendili, Y. *et al*. Raman open database: first interconnected raman–x-ray diffraction open-access resource for material identification. *Journal of applied crystallography* **52**, 618–625 (2019).
10. McMahon, B. The role of journals in maintaining data integrity: checking of crystal structure data in acta crystallographica. *Journal of research of the National Institute of Standards and Technology* **101**, 347 (1996).
11. Brown, I. D. & McMahon, B. CIF: the computer language of crystallography. *Acta Crystallographica Section B: Structural Science* **58**, 317–324 (2002).
12. Murray-Rust, P. & Rzepa, H. Chemical Markup Language, http://www.xml-cml.org, accessed on July 4, 2023 (2012).
13. Murray-Rust, P., Townsend, J. A., Adams, S. E., Phadungsukanan, W. & Thomas, J. The semantics of chemical markup language (CML): dictionaries and conventions. *J. Cheminformatics* **3**, 1–12 (2011).
14. Murray-Rust, P. & Rzepa, H. S. CML: Evolution and design. *J. Cheminformatics* **3**, 44 (2011).
15. Murray-Rust, P. & Rzepa, H. Chemical Markup Language, http://www.xml-cml.org/schema/schema3, accessed on July 4, 2023 (2012).
16. Frisch, M. J. *et al*. Gaussian 03. Gaussian, Inc., Wallingford, CT, http://www.gaussian.com, accessed on July 4, 2023 (2004).
17. Valiev, M. *et al*. NWChem: A comprehensive and scalable open-source solution for large scale molecular simulations. *Comput. Phys. Commun*. **181**, 1477–1489. http://www.nwchem-sw.org/, accessed on July 4, 2023 (2010).

18. An example of a CML mapping from a NWCHEM calculation can be found at: P. Murray-Rust and H. Rzepa, "Chemical Markup Language", http://www.xml-cml.org/examples/schema3/compchem, accessed on July 4, 2023 (2012).
19. McDonald, R. S. & Wilks, P. A. Jr JCAMP-DX: A standard form for exchange of infrared spectra in computer readable form. *Appl. Spectrosc.* **42**, 151–162 (1988).
20. Davies, A. N. & Lampen, P. JCAMP-DX for NMR. *Appl. Spectrosc.* **47**, 1093–1099 (1993).
21. Lampen, P., Hillig, H., Davies, A. N. & Linscheid, M. JCAMP-DX for mass spectrometry. *Appl. Spectrosc.* **48**, 1545–1552 (1994).
22. Baumbach, J. I., Davies, A. N., Lampen, P. & Schmidt, H. JCAMP-DX. A standard format for the exchange of ion mobility spectrometry data (IUPAC recommendations 2001). *Pure Appl. Chem.* **73**, 1765–1782 (2001).
23. Gonze, X. *et al*. Extensible and portable file format for electronic structure and crystallographic data. *Psi-k Newsletters & Highlights* **53** (2007).
24. Gonze, X. *et al*. Specification of an extensible and portable file format for electronic structure and crystallographic data. *Comput. Mater. Sci.* **43**, 1056–1065 (2008).
25. Caliste, D., Pouillon, Y., Verstraete, M. J., Olevano, V. & Gonze, X. Sharing electronic structure and crystallographic data with etsf_io. *Comput. Phys. Commun.* **179**, 748–758 (2008).
26. Ghiringhelli, L. M. *et al*. Towards efficient data exchange and sharing for big-data driven materials science: metadata and data formats. *Npj Comput. Mater.* **3**, 1–9 (2017).
27. Oliveira, M. J. T. *et al*. The CECAM electronic structure library and the modular software development paradigm. *J. Chem. Phys.* **153**, 024117 (2020).
28. Curtarolo, S. *et al*. AFLOWLIB.ORG: A distributed materials properties repository from high-throughput *ab initio* calculations. *Comp. Mat. Sci.* **58**, 227–235, 10.1016/j.commatsci.2012.02.002. http://aflowlib.org, accessed on July 4, 2023 (2012).
29. Talirz, L. *et al*. Materials Cloud, a platform for open computational science. *Sci. Data* **7**, 1–12 (2020).
30. Jain, A. *et al*. The Materials Project: A materials genome approach to accelerating materials innovation. *APL Mater*. **1**, 011002, 10.1063/1.4812323. https://materialsproject.org, accessed on July 4, 2023 (2013).
31. Draxl, C. & Scheffler, M. NOMAD: The FAIR concept for big data-driven materials science. *MRS Bull*. **43**, 676–682. https://nomad-lab.eu, accessed on July 4, 2023 (2018).
32. Draxl, C. & Scheffler, M. The NOMAD laboratory: from data sharing to artificial intelligence. *JPhys Materials* **2**, 036001 (2019).
33. Draxl, C. & Scheffler, M. Big data-driven materials science and its FAIR data infrastructure. In Andreoni, W. & Yip, S. (eds.) *Handbook of Materials Modeling: Methods: Theory and Modeling*, 49–73 (Springer, Cham, 2020).
34. Kirklin, S. *et al*. The open quantum materials database (OQMD): assessing the accuracy of dft formation energies. *Npj Comput. Mater*. **1**, 1–15. http://oqmd.org, accessed on July 4, 2023 (2015).
35. Merkys, A. *et al*. A posteriori metadata from automated provenance tracking: integration of AiiDA and TCOD. *J. Cheminformatics* **9**, 56. http://www.crystallography.net/tcod, accessed on July 4, 2023 (2017).
36. Pizzi, G., Cepellotti, A., Sabatini, R., Marzari, N. & Kozinsky, B. AiiDA: automated interactive infrastructure and database for computational science. *Comput. Mater. Sci.* **111**, 218–230, 10.1016/j.commatsci.2015.09.013. http://www.aiida.net, accessed on July 4, 2023 (2016).
37. Huber, S. P. *et al*. AiiDA 1.0, a scalable computational infrastructure for automated reproducible workflows and data provenance. *Sci. Data* **7**, 1–18 (2020).
38. Uhrin, M., Huber, S. P., Yu, J., Marzari, N. & Pizzi, G. Workflows in AiiDA: Engineering a high-throughput, event-based engine for robust and modular computational workflows. *Comp. Mat. Sci.* **187**, 110086 (2021).
39. Larsen, A. H. *et al*. The atomic simulation environment—a Python library for working with atoms. *J. Condens. Matter Phys*. **29**, 273002. https://wiki.fysik.dtu.dk/ase, accessed on July 4, 2023 (2017).
40. Tadmor, E. B., Elliott, R. S., Sethna, J. P., Miller, R. E. & Becker, C. A. The potential of atomistic simulations and the knowledgebase of interatomic models. *JOM* **63**, 17, 10.1007/s11837-011-0102-6. https://openkim.org, accessed on July 4, 2023 (2011).
41. Andersen, C. W. *et al*. OPTIMADE, an API for exchanging materials data. *Sci. Data* **8**, 217. https://www.optimade.org, accessed on July 4, 2023 (2021).
42. EMMC and EMMO Governance Committee, EMMO - Elementary Multiperspective Material Ontology, https://emmc.info/emmo-info, accessed on July 4, 2023 (2021).
43. Fielding, R. T. *Architectural styles and the design of network-based software architectures* (University of California, Irvine, 2000).
44. The NOMAD team, 2014–2023, https://nomad-lab.eu/prod/v1/gui/search/entries, accessed on July 4, 2023.
45. The NOMAD team, 2015–2023, https://nomad-lab.eu/prod/rae/encyclopedia/, accessed on July 4, 2023.
46. Ghiringhelli, L. M. An AI-toolkit to develop and share research into new materials. *Nat. Rev. Phys.* **3**, 724–724 (2021).
47. Sbailò, L., Fekete, Á., Ghiringhelli, L. M. & Scheffler, M. The NOMAD artificial-intelligence toolkit: turning materials-science data into knowledge and understanding. *npj Computational Materials* **8**, 250 (2022).
48. The NOMAD team, 2018–2023, https://nomad-lab.eu/AIToolkit, accessed on July 4, 2023.
49. Marques, M. A. L., Oliveira, M. J. T. & Burnus, T. Libxc: A library of exchange and correlation functionals for density functional theory. *Comput. Phys. Commun.* **183**, 2272–2281 (2012).
50. Lehtola, S., Steigemann, C., Oliveira, M. J. & Marques, M. A. Recent developments in Libxc — A comprehensive library of functionals for density functional theory. *SoftwareX* **7**, 1–5, https://doi.org/10.1016/j.softx.2017.11.002 (2018).
51. Pritchard, B. P., Altarawy, D., Didier, B., Gibson, T. D. & Windus, T. L. New basis set exchange: An open, up-to-date resource for the molecular sciences community. *Journal of chemical information and modeling* **59**, 4814–4820 (2019).
52. Blum, V. *et al*. Ab initio molecular simulations with numeric atom-centered orbitals. *Comput. Phys. Commun.* **180**, 2175–2196 (2009).
53. Barca, G. M. J. *et al*. Recent developments in the general atomic and molecular electronic structure system. *The Journal of Chemical Physics* **152**, 154102, https://doi.org/10.1063/5.0005188 (2020).
54. Kresse, G. & Furthmüller, J. Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set. *Physical Review B* **54**, 11169–11186 (1996).
55. Marques, M. A., Castro, A., Bertsch, G. F. & Rubio, A. Octopus: a first-principles tool for excited electron–ion dynamics. *Computer Physics Communications* **151**, 60–78 (2003).
56. Segall, M. *et al*. First-principles simulation: ideas, illustrations and the castep code. *Journal of physics: condensed matter* **14**, 2717 (2002).
57. Shang, H., Carbogno, C., Rinke, P. & Scheffler, M. Lattice dynamics calculations based on density-functional perturbation theory in real space, *Comput. Phys. Commun.* **215**, 26 (2017).
58. Giannozzi, P. *et al*. Quantum espresso: a modular and open-source software project for quantum simulations of materials. *Journal of physics: Condensed matter* **21**, 395502 (2009).
59. Gonze, X. *et al*. Abinit: First-principles approach to material and nanosystem properties. *Computer Physics Communications* **180**, 2582–2615 (2009).
60. Lamprecht, A.-L. *et al*. Towards FAIR principles for research software. *Data Science* **3**, 37–59 (2020).
61. Barker, M. *et al*. Introducing the FAIR Principles for research software. *Scientific Data* **9**, 622, https://doi.org/10.1038/s41597-022-01710-x (2022).
62. Katz, D. S. *et al*. Recognizing the value of software: a software citation guide. *F1000Research* **9** (2020).
63. Smith, A. M., Katz, D. S. & Niemeyer, K. E. Software citation principles. *PeerJ Computer Science* **2**, e86 (2016).

64. Terry Bollinger, Terry Bollinger online resources, 2003–2012, http://www.terrybollinger.com/index.html#open_source_reports, accessed on July 4, 2023.
65. Richard Stallman, FLOSS and FOSS, 2021, https://www.gnu.org/philosophy/floss-and-foss.html, accessed on July 4, 2023.
66. Lejaeghere, K. *et al*. Reproducibility in density functional theory calculations of solids. *Science* **351**, aad3000 (2016).
67. Schäfer, T. *et al*. Tracking the footprints of spin fluctuations: A multimethod, multimessenger study of the two-dimensional Hubbard model. *Physical Review X* **11**, 011058 (2021).
68. Zepeda-Ruiz, L. A., Stukowski, A., Oppelstrup, T. & Bulatov, V. V. Probing the limits of metal plasticity with molecular dynamics simulations. *Nature* **550**, 492–495, https://doi.org/10.1038/nature23472 (2017).
69. Miyoshi, E. *et al*. Large-scale phase-field simulation of three-dimensional isotropic grain growth in polycrystalline thin films. *Modell. Simul. Mater. Sci. Eng.* **27**, 054003, https://doi.org/10.1088/1361-651X/ab1e8b (2019).
70. Shibuta, Y. *et al*. Heterogeneity in homogeneous nucleation from billion-atom molecular dynamics simulation of solidification of pure metal. *Nat. Commun.* **8**, 10, https://doi.org/10.1038/s41467-017-00017-5 (2017).
71. Piana, S., Klepeis, J. L. & Shaw, D. E. Assessing the accuracy of physical models used in protein-folding simulations: quantitative evidence from long molecular dynamics simulations. *Curr. Opin. Struct. Biol.* **24**, 98–105, https://doi.org/10.1016/j.sbi.2013.12.006 (2014).
72. Leitherer, A., Ziletti, A. & Ghiringhelli, L. M. Robust recognition and exploratory analysis of crystal structures via Bayesian deep learning. *Nat. Commun.* **12**, 1–13 (2021).
73. Cummings, P. T. *et al*. Open-source molecular modeling software in chemical engineering focusing on the molecular simulation design framework. *AIChE Journal* **67**. https://mosdef.org/, accessed on July 4, 2023 (2021).
74. Laio, A. & Parrinello, M. Escaping free-energy minima. *Proc. Natl. Acad. Sci. USA* **99**, 12562–12566, https://doi.org/10.1073/pnas.202427399 (2002).
75. Torrie, G. M. & Valleau, J. P. Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling. *J. Comput. Phys.* **23**, 187–199, https://doi.org/10.1016/0021-9991(77)90121-8 (1977).
76. Marinari, E. & Parisi, G. Simulated tempering: a new Monte Carlo scheme. *EPL* **19**, 451 (1992).
77. Dellago, C., Bolhuis, P. & Geissler, P. L. Transition path sampling. *Adv. Chem. Phys.* **123** (2002).
78. Allen, R. J., Valeriani, C. & Rein Ten Wolde, P. Forward flux sampling for rare event simulations. *J. Phys. Condens. Matter* **21**, 463102, https://doi.org/10.1088/0953-8984/21/46/463102 (2009).
79. Tribello, G. A., Bonomi, M., Branduardi, D., Camilloni, C. & Bussi, G. PLUMED 2: New feathers for an old bird. *Comput. Phys. Commun.* **185**, 604–613, https://doi.org/10.1016/j.cpc.2013.09.018 (2014).
80. Greff da Silveira, L., Jacobs, M., Prampolini, G., Livotto, P. R. & Cacelli, I. Development and validation of quantum mechanically derived Force-Fields: Thermodynamic, structural, and vibrational properties of aromatic heterocycles. *J. Chem. Theory Comput.* **14**, 4884–4900, https://doi.org/10.1021/acs.jctc.8b00218 (2018).
81. Abraham, M. J. *et al*. GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **1**, 19–25 (2015).
82. Todorov, I. T., Smith, W., Trachenko, K. & Dove, M. T. DL_POLY_3: new dimensions in molecular dynamics simulations via massive parallelism. *Journal of Materials Chemistry* **16**, 1911–1918 (2006).
83. Thompson, A. P. *et al*. LAMMPS-a flexible simulation tool for particle-based materials modeling at the atomic, meso, and continuum scales. *Computer Physics Communications* **271**, 108171 (2022).
84. Kühne, T. D. *et al*. CP2K: An electronic structure and molecular dynamics software package-quickstep: Efficient and accurate electronic structure calculations. *The Journal of Chemical Physics* **152** (2020).
85. Kapil, V. *et al*. i-PI 2.0: A universal force engine for advanced molecular simulations. *Computer Physics Communications* **236**, 214–223 (2019).
86. The PLUMED consortium, https://www.plumed-nest.org/, accessed on July 4, 2023 (2019).
87. The PLUMED consortium. Promoting transparency and reproducibility in enhanced molecular simulations. *Nat. Methods* **16**, 670–673, https://doi.org/10.1038/s41592-019-0506-8 (2019).
88. Brehm, M. & Thomas, M. An efficient lossless compression algorithm for trajectories of atom positions and volumetric data. *J. Chem. Inf. Model.* **58**, 2092–2107, https://doi.org/10.1021/acs.jcim.8b00501 (2018).
89. Deelman, E., Gannon, D., Shields, M. & Taylor, I. Workflows and e-science: An overview of workflow system features and capabilities. *Future Gener. Comput. Syst.* **25**, 528, https://doi.org/10.1016/j.future.2008.06.012 (2009).
90. Curtarolo, S. *et al*. AFLOW: An automatic framework for high-throughput materials discovery. *Comp. Mat. Sci.* **58**, 218–226, https://doi.org/10.1016/j.commatsci.2012.02.005 (2012).
91. Calderon, C. E. *et al*. The AFLOW standard for high-throughput materials science calculations. *Comp. Mat. Sci.* **108**(Part A), 233–238, https://doi.org/10.1016/j.commatsci.2015.07.019 (2015).
92. Pizzi, G. Open-science platform for computational materials science: AiiDA and the Materials Cloud. In Andreoni, W. & Yip, S. (eds.) *Handbook of Materials Modeling: Methods: Theory and Modeling*, 1–24, https://doi.org/10.1007/978-3-319-42913-7_64-1 (Springer International Publishing, Cham, 2018).
93. Jain, A. *et al*. Fireworks: a dynamic workflow system designed for high-throughput applications. *Concurr Comput*. **27**, 5037–5059. https://materialsproject.org, accessed on July 4, 2023 (2015).
94. Bröder, J., Wortmann, D. & Blügel, S. Using the AiiDA-FLEUR package for all-electron ab initio electronic structure data generation and processing in materials science. In Schultz, M., Pleiter, D. & Bauer, P. (eds.) *Extreme Data: Demands, Technologies, and Services Workshop Proceedings*, vol. 40, 43–47 (Forschungszentrum Jülich, 2018).
95. Toher, C. *et al*. High-throughput computational screening of thermal conductivity, Debye temperature, and Grüneisen parameter using a quasiharmonic Debye model. *Phys. Rev. B* **90**, 174107, https://doi.org/10.1103/PhysRevB.90.174107 (2014).
96. Huber, S. P. *et al*. Common workflows for computing material properties using different quantum engines. *npj Computational Materials* **7**, 1–12 (2021).
97. Togo, A. & Tanaka, I. First principles phonon calculations in materials science. *Scr. Mater.* **108**, 1–5 (2015).
98. Petretto, G., Gonze, X., Hautier, G. & Rignanese, G.-M. Convergence and pitfalls of density functional perturbation theory phonons calculations from a high-throughput perspective. *Comput. Mater. Sci.* **144**, 331–337 (2018).
99. Petretto, G. *et al*. High-throughput density-functional perturbation theory phonons for inorganic materials. *Sci. Data* **5**, 180065 EP– (2018).
100. Mounet, N. *et al*. Two-dimensional materials from high-throughput computational exfoliation of experimentally known compounds. *Nat. Nanotechnol.* **13**, 246–252 (2018).
101. Knoop, F., Purcell, T. A. R., Scheffler, M. & Carbogno C., FHI-Vibes: *Ab initio* vibrational simulations, *J. Open Source Softw.* **5**, 2671 (2020)
102. de Jong, M. *et al*. Charting the complete elastic properties of inorganic crystalline compounds. *Sci. Data* **2**, 150009, https://doi.org/10.1038/sdata.2015.9 (2015).
103. Toher, C. *et al*. Combining the AFLOW GIBBS and elastic libraries to efficiently and robustly screen thermomechanical properties of solids. *Phys. Rev. Mater.* **1**, 015401, https://doi.org/10.1103/PhysRevMaterials.1.015401 (2017).
104. Golesorkhtabar, R., Pavone, P., Spitaler, J., Puschnig, P. & Draxl, C. Elastic: A tool for calculating second-order elastic constants from first principles. *Comput. Phys. Commun.* **184**, 1861–1873 (2013).

105. Oses, C., Toher, C. & Curtarolo, S. Data-driven design of inorganic materials with the automatic flow framework for materials discovery. *MRS Bull.* **43**, 670–675, https://doi.org/10.1557/mrs.2018.207 (2018).
106. Plata, J. J. *et al*. An efficient and accurate framework for calculating lattice thermal conductivity of solids: AFLOW-AAPL Automatic Anharmonic Phonon Library. *npj Comput. Mater.* **3**, 45, https://doi.org/10.1038/s41524-017-0046-7 (2017).
107. Yang, K., Oses, C. & Curtarolo, S. Modeling off-stoichiometry materials with a high-throughput *ab-initio* approach. *Chem. Mater.* **28**, 6484–6492, https://doi.org/10.1021/acs.chemmater.6b01449 (2016).
108. Symalla, F. *et al*. Charge Transport by Superexchange in Molecular Host-Guest Systems. *Phys. Rev. Lett.* **117**, 276803–6 (2016).
109. Friederich, P. *et al*. Rational In Silico Design of an Organic Semiconductor with Improved Electron Mobility. *Adv. Mater.* **29**, 1703505–7 (2017).
110. Mercado, R. *et al*. In silico design of 2d and 3d covalent organic frameworks for methane storage applications. *Chem. Mater.* **30**, 5069–5086 (2018).
111. Rüßmann, P., Bertoldo, F. & Blügel, S. The AiiDA-KKR plugin and its application to high-throughput impurity embedding into a topological insulator. *Npj Comput. Mater.* **7**, 1–9 (2021).
112. Vitale, V. *et al*. Automated high-throughput wannierisation. *Npj Comput. Mater*. **6** (2020).
113. Mishra, S. *et al*. Observation of fractional edge excitations in nanographene spin chains. *Nature* **598**, 287–292 (2021).
114. Folk, M., Heber, G., Koziol, Q., Pourmal, E. & Robinson, D. An overview of the hdf5 technology suite and its applications. In *Proceedings of the EDBT/ICDT 2011 workshop on array databases*, 36–47 (2011).
115. Könnecke, M. *et al*. The NeXus data format. *Journal of applied crystallography* **48**, 301–305. https://www.nexusformat.org/, accessed on July 4, 2023 (2015).
116. Delagenière, S. *et al*. ISPyB: an information management system for synchrotron macromolecular crystallography. *Bioinformatics* **27**, 3186–3192 (2011).
117. Malbet-Monaco, S., Leonard, G. A., Mitchell, E. P. & Gordon, E. J. How the ESRF helps industry and how they help the ESRF. *Acta Crystallogr. D* **69**, 1289–1296 (2013).
118. Fisher, S., Levik, K., Williams, M., Ashton, A. & McAuley, K. SynchWeb: a modern interface for ISPyB. *J. Appl. Crystallogr.* **48**, 927–932 (2015).
119. De Maria Antolinos, A. *et al*. ISPyB for BioSAXS, the gateway to user autonomy in solution scattering experiments. *Acta Crystallogr. D* **71**, 76–85 (2015).
120. Carpi, N., Minges, A. & Piel, M. eLabFTW: An open source laboratory notebook for research labs. *J. Open Source Softw.* **2**, 146 (2017).
121. Bricogne, G. *et al*. Achieving higher performance in high-throughput compound and fragment screening campaigns by the use of" club class" data collection with pipedream and crims. *Acta Crystallogr. A* **74**, A248–A248 (2018).
122. Tremouilhac, P. *et al*. Chemotion ELN: an open source electronic lab notebook for chemists in academia. *J. Cheminformatics* **9**, 1–13 (2017).
123. Ghedini, E., Friis, J., Goldbeck, G., Prinz, M. & Bleken, F. 2019–2022, https://github.com/emmo-repo/EMMO, accessed on July 4, 2023.
124. Ashino, T. Materials ontology: An infrastructure for exchanging materials information and knowledge. *Data Sci. J.* **9**, 54–61 (2010).
125. Cheung, K., Drennan, J. & Hunter, J. Towards an ontology for data-driven discovery of new materials. In *AAAI Spring Symposium: Semantic Scientific Knowledge Integration*, 9–14 (2008).
126. Thomas, D. G., Pappu, R. V. & Baker, N. A. Nanoparticle ontology for cancer nanotechnology research. *J. Biomed. Inform* **44**, 59–74 (2011).
127. Hastings, J. *et al*. enanomapper: harnessing ontologies to enable data integration for nanomaterial risk assessment. *J. Biomed. Semantics* **6**, 1–15 (2015).
128. Li, H., Armiento, R. & Lambrix, P. An ontology for the materials design domain. In *International Semantic Web Conference*, 212–227 (Springer, 2020).
129. Li, H., Armiento, R. & Lambrix, P. A method for extending ontologies with application to the materials science domain. *Data Sci. J.* **18**, 1–21 (2019).
130. Himmer-Lenz, M.-O. PhD thesis, Humboldt Universität zu Berlin https://doi.org/10.18452/24340 (2022).
131. Kuban, M., Rigamonti, S., Scheidgen, M. & Draxl, C. Density-of-states similarity descriptor for unsupervised learning from materials data. *Scientific Data* **9**, 646 (2022).
132. Kuban, M. *et al*. Similarity of materials and data-quality assessment by fingerprinting. *MRS Bulletin* **47**, 991–999 (2022).
133. Psi-k Network, 1994–2023, https://psi-k.net, accessed on July 4, 2023.
134. Olivetti, E. A. *et al*. Data-driven materials research enabled by natural language processing and information extraction. *Appl.Phys. Rev.* **7**, 041317 (2020).

Acknowledgements

We would like to thank all the participants to the workshop "Shared Metadata and Data Formats for Big-Data Driven Materials Science: A NOMAD–FAIR-DI Workshop", as listed at META2019, who have contributed with questions and comments to ideas discussed in this paper. The organizers of and participants to the OMDI2021 workshop (see https://liu.se/en/research/omdi2021 for the full list of names) are acknowledged for insightful discussions that inspired some of the concepts discussed in Section "Outlook on ontologies in materials science". This work received funding by the European Union's Horizon 2020 research and innovation program under the grant agreement N° 951786 (NOMAD CoE) and by the German Research Foundation (DFG) through the NFDI consortium FAIRmat, project 460197019. We acknowledge support by the Open Access Publication Fund of Humboldt-Universität zu Berlin. SVL's contribution was supported by RSCF grant 21-13-00419.

Author contributions

The present paper is inspired by and based on the minutes of the work-groups discussions at the workshop "Shared Metadata and Data Formats for Big-Data Driven Materials Science: A NOMAD–FAIR-DI Workshop". Here, we report the composition of the original work groups, which reflect into the main contributions to the paper's sections. Metadata, metadata schemas and ontologies (Introduction, Section "Towards FAIR metadata schemas for computational materials science" and section "Outlook on ontologies in materials science"): Patrick Lambrix, Javad Chamanara, Carsten Baldauf, Tatyana Sheveleva, Benjamin Regler, Alvin Noe Ladines, Christoph T. Koch, Christof Wöll, Stefano Cozzini, Astrid Schneidewind, Maja-Olivia Himmer; Ground-state calculations (Section "Metadata for ground-state electronic-structure calculations"): Micael Oliveira, Sergey Levchenko; Perturbative and excited-states calculations (Section "Metadata for external-perturbation and excited-state electronic-structure calculations"): Claudia Draxl, Pasquale Pavone, Denis Usvyat; Potential-energy sampling (Section "Metadata for potential-energy sampling"): James Kermode, Tristan Bereau, Christian Carbogno, Omar Valsson, Markus Kühbach, Chuanxun Su, Ron Miller, Berk Onat; Workflows (Section "Metadata for Computational Workflows"): Stefano Curtarolo, Shyam Dwaraknath, Adam Michalchuk, Giovanni Pizzi, Gian-

Marco Rignanese, Jörg Schaarschmidt; Data formats (Section "File Formats"): Ádám Fekete, Markus Scheidgen; Metadata for experiments (Section "Metadata schemas for experimental materials science"): Christoph T. Koch, Sandor Brockhauser, Astrid Schneidewind. Luca M. Ghiringhelli and Matthias Scheffler coordinated the formation of the work groups, participated to the discussions in several work groups, and prepared the first draft of the paper. All authors contributed to the final version of the paper.

Funding
Open Access funding enabled and organized by Projekt DEAL.

Competing interests
The authors declare no competing interests.

Additional information
Correspondence and requests for materials should be addressed to L.M.G.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Luca M. Ghiringhelli[1,2,25]✉, Carsten Baldauf[3], Tristan Bereau[4], Sandor Brockhauser[1], Christian Carbogno[2], Javad Chamanara[5], Stefano Cozzini[6], Stefano Curtarolo[7], Claudia Draxl[1,2], Shyam Dwaraknath[8], Ádám Fekete[1], James Kermode[9], Christoph T. Koch[1], Markus Kühbach[1], Alvin Noe Ladines[1], Patrick Lambrix[10], Maja-Olivia Himmer[2], Sergey V. Levchenko[11], Micael Oliveira[12], Adam Michalchuk[13,14], Ronald E. Miller[15], Berk Onat[9], Pasquale Pavone[1], Giovanni Pizzi[16,17], Benjamin Regler[2], Gian-Marco Rignanese[18], Jörg Schaarschmidt[19], Markus Scheidgen[1], Astrid Schneidewind[20], Tatyana Sheveleva[5], Chuanxun Su[21], Denis Usvyat[22], Omar Valsson[23], Christof Wöll[24] & Matthias Scheffler[1,2]

[1]Physics Department and IRIS Adlershof, Humboldt-Universität zu Berlin, Berlin, Germany. [2]The NOMAD Laboratory at the Fritz-Haber-Institut of the Max-Planck-Gesellschaft and IRIS-Adlershof of the Humboldt-Universität zu Berlin, Berlin, Germany. [3]Fritz-Haber-Institut of the Max-Planck-Gesellschaft, Berlin, Germany. [4]Van't Hoff Institute for Molecular Sciences and Informatics Institute, University of Amsterdam, Amsterdam, 1098 XH, The Netherlands. [5]TIB – Leibniz Information Centre for Science and Technology and University Library, 30167, Hanover, Germany. [6]AREA Science Park, località Padriciano, 34149, Trieste, Italy. [7]Center for Autonomous Materials Design and Department of Mechanical Engineering and Materials Science, Duke University, Durham, NC, 27708, USA. [8]Lawrence Berkeley National Laboratory, Berkeley, CA, USA. [9]Warwick Centre for Predictive Modelling, School of Engineering, University of Warwick, Coventry, CV4 7AL, United Kingdom. [10]Department of Computer and Information Science and The Swedish e-Science Research Centre, Linköping University, Linköping, Sweden. [11]Center for Energy Science and Technology, Skolkovo Institute of Science and Technology, Moscow, Russia. [12]Max Planck Institute for the Structure and Dynamics of Matter, Hamburg, Germany. [13]Federal Institute for Materials Research and Testing (BAM), 12489, Berlin, Germany. [14]School of Chemistry, University of Birmingham, B15 2TT, Edgbaston, Birmingham, UK. [15]Department of Mechanical and Aerospace Engineering, Carleton University, Ottawa, ON, K1S 5B6, Canada. [16]Theory and Simulation of Materials (THEOS) and National Centre for Computational Design and Discovery of Novel Materials (MARVEL), École Polytechnique Fédérale de Lausanne, CH-1015, Lausanne, Switzerland. [17]Laboratory for Materials Simulations (LMS), Paul Scherrer Institut (PSI), CH-5232, Villigen, Switzerland. [18]Institute of Condensed Matter and Nanosciences (IMCN), UCLouvain, Chemin des Étoiles 8, B-1348, Louvain-la-Neuve, Belgium. [19]Institute of Nanotechnology, Karlsruhe Institute of Technology (KIT), 76344 Eggenstein-Leopoldshafen, Karlsruhe, Germany. [20]Jülich Center for Neutron Science at MLZ, Forschungszentrum Jülich GmbH, Lichtenbergstrase 1, 85748, Garching, Germany. [21]CAS Key Laboratory of Quantum Information, University of Science and Technology of China, Hefei, 230026, People's Republic of China. [22]Chemistry Department, Humboldt-Universität zu Berlin, Berlin, Germany. [23]Department of Chemistry, University of North Texas, Denton, TX, 76201, USA. [24]Institute of Functional Interfaces, Karlsruhe Institute of Technology (KIT), 76344 Eggenstein-Leopoldshafen, Karlsruhe, Germany. [25]Present address: Department of Materials Science and Engineering, Friedrich-Alexander Universität, Erlangen-Nürnberg, Germany. ✉e-mail: luca.ghiringhelli@physik.hu-berlin.de