Perspective

FAIR data enabling new horizons for materials research

https://doi.org/10.1038/s41586-022-04501-x

Received: 8 March 2021

Accepted: 28 January 2022

Published online: 27 April 2022

Check for updates

Matthias Scheffler[1,2], Martin Aeschlimann[3], Martin Albrecht[4], Tristan Bereau[5], Hans-Joachim Bungartz[6], Claudia Felser[7], Mark Greiner[8], Axel Groß[9], Christoph T. Koch[1], Kurt Kremer[5], Wolfgang E. Nagel[10], Markus Scheidgen[1], Christof Wöll[11] & Claudia Draxl[1,2]✉

The prosperity and lifestyle of our society are very much governed by achievements in condensed matter physics, chemistry and materials science, because new products for sectors such as energy, the environment, health, mobility and information technology (IT) rely largely on improved or even new materials. Examples include solid-state lighting, touchscreens, batteries, implants, drug delivery and many more. The enormous amount of research data produced every day in these fields represents a gold mine of the twenty-first century. This gold mine is, however, of little value if these data are not comprehensively characterized and made available. How can we refine this feedstock; that is, turn data into knowledge and value? For this, a FAIR (findable, accessible, interoperable and reusable) data infrastructure is a must. Only then can data be readily shared and explored using data analytics and artificial intelligence (AI) methods. Making data 'findable and AI ready' (a forward-looking interpretation of the acronym) will change the way in which science is carried out today. In this Perspective, we discuss how we can prepare to make this happen for the field of materials science.

The number of possible materials is practically infinite. But even for the so-far known materials, our knowledge about their properties and their synthesis is very shallow. There is no doubt that forms of condensed matter exist, or can be created, that show better, or even new, properties and functions than the materials that are known and used today. How can we find them? High-throughput screening of materials – experimentally or theoretically – collects important information. These results will boost new discoveries, but the immensity of possible materials cannot be covered by such explicit searches. Moreover, in current purpose-focused research, only a small fraction of the data produced in the studies is published, and many data are not fully characterized. Furthermore, the metadata (the information that explains and characterizes the measured or calculated data), ontologies (the relationships in metadata) and workflows of different research groups cannot be easily reconciled. Thus, most research data are neither findable nor interoperable.

A FAIR data infrastructure will foster the exchange of scientific information. The meaning of the acronym—that data should be findable, accessible, interoperable and reusable—is explained in the original publication by Wilkinson et al.[1] and elaborated on, for example, at the GO FAIR web pages (https://go-fair.org/fair-principles/). The crucial and very laborious first step towards the FAIRification of data concerns the need to comprehensively describe data by metadata; that is, to characterize data fully and unambiguously so that the research is reproducible. Then scientists, engineers and others can also combine data and metadata from different studies and use them in different contexts. This will open synergies between materials science subdomains and facilitate inter-institute and cross-discipline research. It will also enable data to be used for deeper analyses and for training AI models. Clearly, a FAIR data infrastructure will also show data provenance.

The US Materials Genome Initiative (MGI, https://mgi.gov/) was announced in 2011 for "discovering, manufacturing, and deploying advanced materials twice as fast and at a fraction of the cost compared to traditional methods." It markedly boosted collaborations and high-throughput experiments and computations. FAIRmat (https://FAIRmat-NFDI.eu/) develops the original MGI concept further by implementing a FAIR data infrastructure for condensed-matter physics and the chemical physics of solids. It is a consortium of the German National Research Data Infrastructure programme (https://nfdi.de). FAIRmat interweaves data and tools from and for materials synthesis, experiments, theory and computation, and makes all data available to the whole materials science community and beyond. In this endeavour, it unites researchers from condensed matter physics, the chemical physics of solids and computer science and IT experts.

Materials science is strongly affected by all the four Vs (4V) of big data: volume (the amount of data), variety (the heterogeneity of form and meaning of data), velocity (the rate at which data may change or new data arrive) and veracity (the uncertainty of data quality). The various experimental and theoretical examples provided below will illustrate these different aspects. In general, a FAIR data infrastructure

[1] Physics Department and IRIS Adlershof, Humboldt-Universität zu Berlin, Berlin, Germany. [2]The NOMAD Laboratory at the Fritz Haber Institute of the Max Planck Society, Berlin, Germany. [3]Department of Physics and Research Center OPTIMAS, University of Kaiserslautern, Kaiserslautern, Germany. [4]Leibniz-Institut für Kristallzüchtung, Berlin, Germany. [5]Max-Planck-Institut für Polymerforschung, Mainz, Germany. [6]Department of Informatics, Technical University of Munich, Munich, Germany. [7]Max Planck Institute for Chemical Physics of Solids, Dresden, Germany. [8]Max Planck Institute for Chemical Energy Conversion, Mülheim an der Ruhr, Germany. [9]Institute of Theoretical Chemistry, Ulm University and Helmholtz-Institute Ulm, Ulm, Germany. [10]Computer Science Department, Technical University Dresden, Dresden, Germany. [11]Institute of Functional Interfaces, Karlsruhe Institute of Technology, Karlsruhe, Germany. ✉e-mail: claudia.draxl@physik.hu-berlin.de

requires an in-depth description of how the data have been obtained, addressing metadata, ontologies and workflows. Obviously, only the experts (those creating the samples or computer codes and performing measurements or calculations that is, producing the data) have the insight and knowledge to provide this critical information.

The topic of this Perspective, as outlined above, includes the request for a notable change in scientific culture. Thinking beyond our present research focus on effect, phenomenon or application requires us to accept publishing 'clean data'—that is, well-characterized and clearly annotated data—represents a value of similar importance to a standard publication, or even higher. This concept carries analogies to Tycho Brahe, who created the data that enabled Johannes Kepler to find his equations and finally led Newton to formulate his theory of gravitation.

Eventually, after having installed an efficient, FAIR research data infrastructure, hosting all data from synthesis, experimental and theoretical studies for a wide range of materials, we also need to pave the way for carrying out new research. Our scientific vision is to build maps of material properties that will guide us in designing and finding new materials for a desired function. This concept follows the spirit of the creation of the periodic table of elements; organizing the roughly 60 atoms known at the time enabled Mendeleev to predict the existence and properties of yet-to-be discovered elements.

In the following, we will describe the state of the art, highlight the challenges and put forward FAIRmat's envisaged solutions.

Data-centric materials science

Science is and always has been based on data, but the term 'data-centric' indicates a radical shift in how information is handled and research is performed. It refers to extensive data collections, digital repositories and new concepts and methods of data analytics. It also implies that we complement traditional purpose-oriented research by using data from other studies.

Some progress in this direction has been made in recent years in terms of collecting data from the many research groups across the planet (all the data, not just what is published in research manuscripts) and making the data FAIR[1,2]. This should be good scientific practice in any case[3,4]. Since 1965, data repositories in materials science have moved towards digitization. A comprehensive list can be found in ref. [5]. Among them, the NOMAD (Novel Materials Discovery) Laboratory (a database for computational materials science; online since 2014, https://nomad-lab.eu/) is unique as it accepts data from practically all computational materials science codes. As it provides the blueprint for FAIRmat, we will summarize its basic concept (for details, see refs. [4,6]). A key guideline of NOMAD (and FAIRmat) is to help scientists and students to upload and download data in the easiest way. In simple terms, data stored at NOMAD are treated analogously to publications at a journal archive, such as https://arxiv.org/. Unlike journal archives, an embargo period can be used for collaborations with selected colleagues or may even be crucial for collaborations with industry. At the time of writing (August 2021), NOMAD contains results from more than 100 million open-access calculations. These are from individual researchers all over the world and include entries from other computational materials databases, such as AFLOW (http://aflow.org), the Materials Project (https://materialsproject.org) and OQMD (the Open Quantum Materials Database, http://oqmd.org). NOMAD converts the data into a common form and provides an easy materials view presentation by means of the NOMAD Encyclopedia (https://nomad-lab.eu/encyclopedia). This allows users to see, compare, explore and understand computed materials data. Furthermore, the NOMAD Artificial Intelligence Toolkit (https://nomad-lab.eu/AIToolkit) offers tools for data analytics and predictions.

The overall challenges of FAIRmat are sketched in Fig. 1: besides organizing and – equally importantly – convincing the community (top left), a critical task concerns the development of metadata standards and ontologies (top right). At present, in materials science, such standards are either totally missing or incomplete. Numerous attempts from standards organizations, such as the International Standards Organization (https://iso.org/), to provide controlled vocabularies, standards for data formats and data handling, have so far failed to reach community-wide adoption.

FAIRmat has already started to establish metadata and dictionaries for digital translations of the vocabulary used in different domains. The next step concerns the description of relations between them, hence, the development of ontologies. They will become particularly important when involved workflows are needed. The NOMAD Meta Info[7] (https://nomad-lab.eu/metainfo) stores descriptive and structured information about materials science data and some interdependencies. Thus, it represents an ontology precursor. There are a lot of discussions regarding ontology within the community; see, for example, refs. [7–9] and the metadata and ontology activities at NIST (https://data.nist.gov/od/dm/nerdm/) and the Materials Ontologies RDA Task Group (https://rd-alliance.org). This also concerns collaborations of FAIRmat with EMMC (https://emmc.info), OPTIMADE[9] (https://optimade.org) and NIST (https://data.nist.gov/).

As illustrated in Fig. 1, data-centric materials science requires a complex infrastructure (bottom right). Established standards for data models in materials science will be considered; for example, CIF (Crystallographic Information Framework, https://iucr.org/resources/cif), CSMD (Core Scientific Metadata Model, http://icatproject-contrib.github.io/CSMD) and NeXus (https://nexusformat.org/). Last but not least, acceptance by researchers requires that the infrastructure also offer support and efficient tools for data processing and analysis (Fig. 1, bottom left).

Other research fields are facing different yet analogous challenges. International contacts, coordination and collaborations of the various fields are promoted by the GO FAIR initiative (https://go-fair.org/), the Research Data Alliance (RDA, https://rd-alliance.org/), the association FAIR-DI (https://fair-di.eu), CODATA (https://codata.org/) and others. A recent publication[10] by Wittenburg et al. on 'FAIR practices in Europe' describes the situation in the areas of humanities, environmental sciences and natural sciences. Although basic concepts and IT tasks are being discussed, true collaborations and reaching the final goal of growing together still need time.

Preparing the research of tomorrow

Putting what is outlined above into practice is a rocky road. To motivate the community to join a culture of extensive data sharing, FAIRmat's policy is to lead by example. Two issues are obviously important to speed up the process and trigger active support: (1) successful, living examples of daily data-centric research[11] to demonstrate what and how things work; and (2) outreach to the wider community, including the education of future scientists and engineers.

To cope with the first point, FAIRmat will demonstrate its approach with specific examples from diverse research fields, including battery research, heterogeneous catalysis, optoelectronics, magnetism and spintronics, multifunctional materials and biophysics. In all of this, FAIRmat will demonstrate the synergistic interplay of materials synthesis, sample preparation and experiments, as well as theory and computation, and provide a much more comprehensive picture than the single subcommunities can achieve. As such, FAIRmat will bring together not only data and tools but, most notably, also people, who will learn each other's 'language'. In fact, the necessary width of competences goes along with a diversity in the nomenclature, which can hamper communication as well as the definition of metadata and ontologies. Likewise, electronic lab notebooks (ELNs) must be standardized to allow seamless integration of data into automatic workflows. Dedicated data-analysis and AI tools will be developed and demonstrated that help to identify the key descriptive physicochemical parameters[12–15]. This will allow for

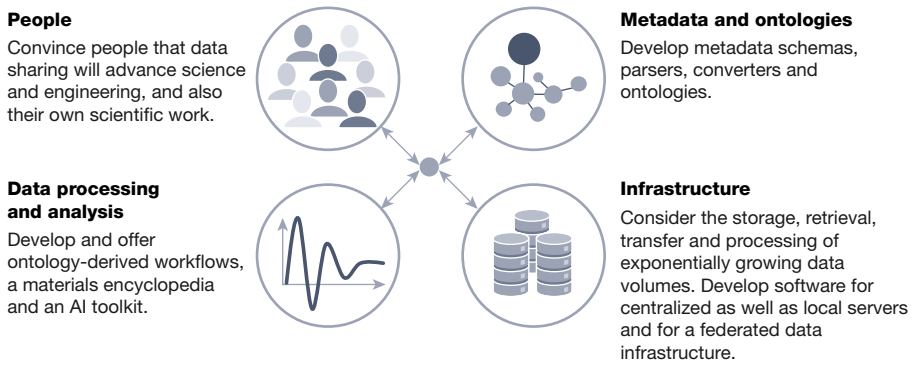

Fig. 1 | Challenges addressed by the FAIRmat initiative. This schematic summarizes the four main categories of challenge faced in implementing FAIRmat.

predictions that go beyond the immediately studied systems and will show trends and enable the identification of materials with statistically exceptional properties[16]. Combining data from different repositories opens further opportunities.

Let us exemplify with two emerging classes of materials that the exploitation of an efficient data infrastructure will be not only helpful but simply mandatory for the digitization of materials research[17]. These examples are high-entropy alloys (HEAs) and metal–organic frameworks (MOFs). For these classes, the sheer number of possible materials is so large that conventional approaches will never be able to unleash even a small part of their full potential. For HEAs, a number of 10^9 possible composite materials with distinctly different properties has been estimated[18], with many of them showing, for example, mechanical properties that exceed by far those of conventional alloys. This huge space of materials further contains HEA oxides with interesting properties in catalysis and energy storage. In the case of MOFs, the situation is even more pronounced. As a result of the huge diversity of MOF building blocks, inorganic clusters and multitopic molecules, the number of compounds is unlimited. Even if one limits the building block weight to that of fullerene (C_{60}), synthesizing only one replica of each compound would already need more atoms than are available on planet Earth. Using AI to analyse the huge amount of experimental information (data for about 100,000 MOFs are stored in databases[19]), we will be able to identify or predict MOFs with particular properties dictated by conceived applications[20]; for example, in optoelectronics[21], biomedicine or catalysis[22].

Turning to the second point—to foster awareness of the importance of FAIR scientific data management and stewardship[1]—FAIRmat will reach out to current students of physics, chemistry, materials science and engineering. We aim to educate a new generation of interdisciplinary researchers, offering classes and lab courses, and to introduce new curricula. A necessary requirement is to convince teachers, professors and other decision makers. The FAIRmat consortium will initiate and organize focused, crosscutting workshops together with, for example, colleagues from chemistry and biochemistry, astroparticle and elementary particle physics, mathematics and engineering. Some topics may be general, such as ontologies or data infrastructure, whereas others will be more specific, including particular experimental techniques or specific simulation methods. Hands-on training, schools and hackathons, as well as the usual online tutorials, will be part of our portfolio. Listening to the needs of small communities or groups will make sure that no one is left behind.

Although industry is very interested in the availability of data, the materials encyclopedia and the AI tools, most investigators hesitate to contribute their own data. Understandably, a company can survive only if they create products that are better or cheaper than those of their competitors. FAIRmat accepts these worries, for example, by allowing for an embargo of uploaded data (see above). The NOMAD Oasis (see also below), which is a key element of the federated FAIRmat infrastructure, can also be operated behind industrial firewalls as a stand-alone server with full functionality.

Science is an international, open activity. So, clearly, all the concepts and plans are and will be discussed, coordinated and implemented together with our colleagues worldwide. In fact, the first FAIR-DI Conference on a FAIR Data Infrastructure for Materials Genomics had 539 participants from all over the world (https://th.fhi-berlin.mpg.de/meetings/fairdi2020/).

Let us end this section by noting that individual researchers already profit from the data infrastructure, even though we are at an early stage in progressing towards the next level of research. For example, countless CPU hours are being saved because computational results are well documented and accessible and do not need to be repeated. Consequently, human time is saved as well and scientists can concentrate on new studies. Students learn faster as they can access extensive reference data. Error or uncertainty estimates are possible and more robust when using well-documented databases. Further results not documented in publications are available in the uploaded data. Studies that were designed for a specific target can now be used for a different topic (repurposing). After receiving a digital object identifier (DOI), uploaded data become citable. This also applies to analytics tools. Although the full potential of FAIRmat will require a larger community to realize and join, the spirit of findable and AI-ready research data has already attracted substantial attention.

FAIR data infrastructure for materials science

FAIRmat will build a federated infrastructure of many domain-specific data-repository solutions: as few as possible but as many as are needed. In NOMAD[6], such individual repositories are called 'oases' and support the different users' local, domain-specific, individual needs to acquire, manage and analyse their data. An oasis is a stand-alone service typically connected to a central server, called the portal, but can also be run independently. As such, it is being tested at present as a building block of FAIRmat's federated data infrastructure.

All participating groups or institutions will manage their data using the FAIRmat frame, a common compute, management and storage concept, with a central metadata repository. To enable 4V data processes, 'federated data with centralized metadata' will be the general principle. Selected data may also be stored centrally, if it is functionally beneficial for users or increases the availability of high-value datasets (see Fig. 2).

The portal will be the gateway for users to access all materials science data. Although popular search engines such as Google search for phrases in generic and mostly text-based properties (domain agnostic), we need to search for precise criteria in materials-science-specific

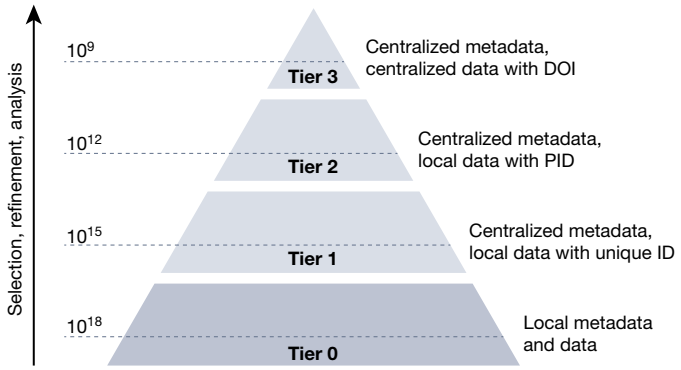

Fig. 2 | Tiers of metadata and data and their approximate volume. The left axis is the volume in bytes. As part of the research process, acquired data are filtered, analysed and refined. This generally produces smaller and smaller datasets of higher value. Owing to the expected data volume, the same level of availability for all data cannot be guaranteed. Therefore, data are categorized into four tiers with different levels of availability, from not shared at all (tier 0) towards published with a DOI (tier 3).

metadata with their individual scientific notations and semantics. Thus, FAIRmat searches are domain-aware.

We will implement a common schema for all FAIRmat metadata and data. However, the data properties that are available for a given type of data differ from method to method and from domain to domain. There will be subsets of common properties for each subdomain, and these subsets form a hierarchy. For example, experiments and synthesis share a common notion of material, measurement or sample. This includes tagging samples with RFID (radio-frequency identification) or QR (quick response) code labels that are linked to every dataset acquired from them. On top of this hierarchy, and even outside the materials science domain, we will always have Dublin Core-style (https://dublincore.org/) metadata about who, where and when.

This bottom-up harmonizing of metadata from different subdomains requires the development of data converters and a shared data schema. This will provide more flexibility when connecting many laboratories and new subdomains than top-down forced adoption of a new data format.

This hierarchy of common properties will also form the basis for exploring all materials science data. Similar to an online shop that allows customers to browse different categories of product, with varying criteria depending on the type of product, the central user interface will allow one to browse different subdomains of materials science on the basis of varying availabilities of data properties. On top of this, one may specify general properties, such as a material's chemical composition and a scientific method. Then, more criteria will be made available. In this way we will design a common encyclopedia that supports the specific needs of the various materials science subdomains but will also provide more general information to non-experts.

Offering convenient tools for data analysis is an overall goal of FAIRmat. An example is the NOMAD Artificial Intelligence Toolkit (https://nomad-lab.eu/AIToolkit). At present, it provides several Jupyter Notebooks, some of them associated with a publication. It is recommended that researchers publish their AI analysis as well or modify or advance existing notebooks for their studies. Uploaded notebooks can obtain a DOI so that they are citable. As some data files will be huge, and may be distributed across several servers and cities, the analysis software will use the centralized metadata and extract the needed information from the (huge) data files. For the latter, we will bring the software to the data, avoiding the transfer of large files.

Other critical issues are long-term and 24/7 data availability (especially in a federated network), safety and security (especially when dealing with published versus unpublished data), data lifecycle (for example, from raw instrument readings to fully analysed and published datasets), linking data between domains, annotating data with a common user identity (for example, through ORCID; https://orcid.org) and more.

FAIR, reproducible synthesis

Synthesizing materials with well-defined properties in a reproducible fashion is of utmost importance to materials science. Unfortunately, this desire is not always fulfilled because it requires control of a large number of experimental details, and the full entirety of the relevant parameters is typically not known. The concept of data-centric science and the development of AI tools promise to model synthesis more reliably and to identify the relevant set of descriptive parameters and their mutual interdependencies, or at least their correlations. Linking synthesis data to data from experimental materials science and theory using common metadata schemas and ontologies will create a new level of the science of materials synthesis.

Publicly accessible databases such as Landolt-Börnstein/Springer Materials (https://materials.springer.com) and the Inorganic Crystal Structure Database (https://icsd.products.fiz-karlsruhe.de) contain huge numbers of entries on the properties of crystalline materials but they lack information on their synthesis. Recently, work on the basis of machine-learning and natural-language-processing techniques has started to codify materials synthesis conditions and parameters that are published in journal articles[23]. The auto-generated open-source dataset at Berkeley (https://ceder.berkeley.edu/text-mined-synthesis) consisted of 19,744 chemical reactions retrieved from 53,538 solid-state synthesis paragraphs by March 2021. However, typically, this information is incomplete, and published information is biased towards reports of successful studies, omitting failed attempts.

This unsatisfactory situation is rooted in the complexity of the synthesis processes, including elaborate workflows and a large diversity of instruments for characterization. In the realm of FAIRmat, we follow the relevant phase transformations that occur during synthesis from the melt, from the gas phase, from solid phases and from solution. Synthesis by assembly complements these classical approaches. The nature of the assembly method is very different, as collective behaviour gives rise to new properties, such as the formation of aggregates or self-assembly. Figure 3 depicts the variety of crystal growth methods. Even though Czochralski, Bridgman, metal flux growth and optical floating zone are all melt-growth techniques—that is, they belong to the same type of phase transition—they are distinguished by the contact of the melt with the crucible, by the seeding of the single crystal and by thermal gradients, with great influence on crystallinity and impurity content. But even fine details matter. For example, the geometry of the reactors, fluctuations in the impurity content of the source material, the flow of precursors in the reactor or the miscut and pretreatment of substrates in epitaxial growth may have detrimental effects. At this point, synthesis is often based on experience and tricks, which are not readily shared with others. Obviously, this makes the development of metadata schemas and ontologies a formidable task and—with respect to the four Vs—synthesis struggles mainly with variety and veracity.

We started to establish metadata and ontologies following the above-mentioned phase transformations. To connect to the other experimental disciplines (for example, sample characterization) we aim at a common ELN scheme and laboratory information management system (LIMS) and uniquely identify the samples, as noted in the infrastructure section. Thereby, we link the measured physical and chemical properties of a specimen to the synthesis workflow. The ELN and LIMS data are automatically fed into a prototype repository that is now being developed at the Leibniz Institute for Crystal Growth (https://ikz-berlin.de/).

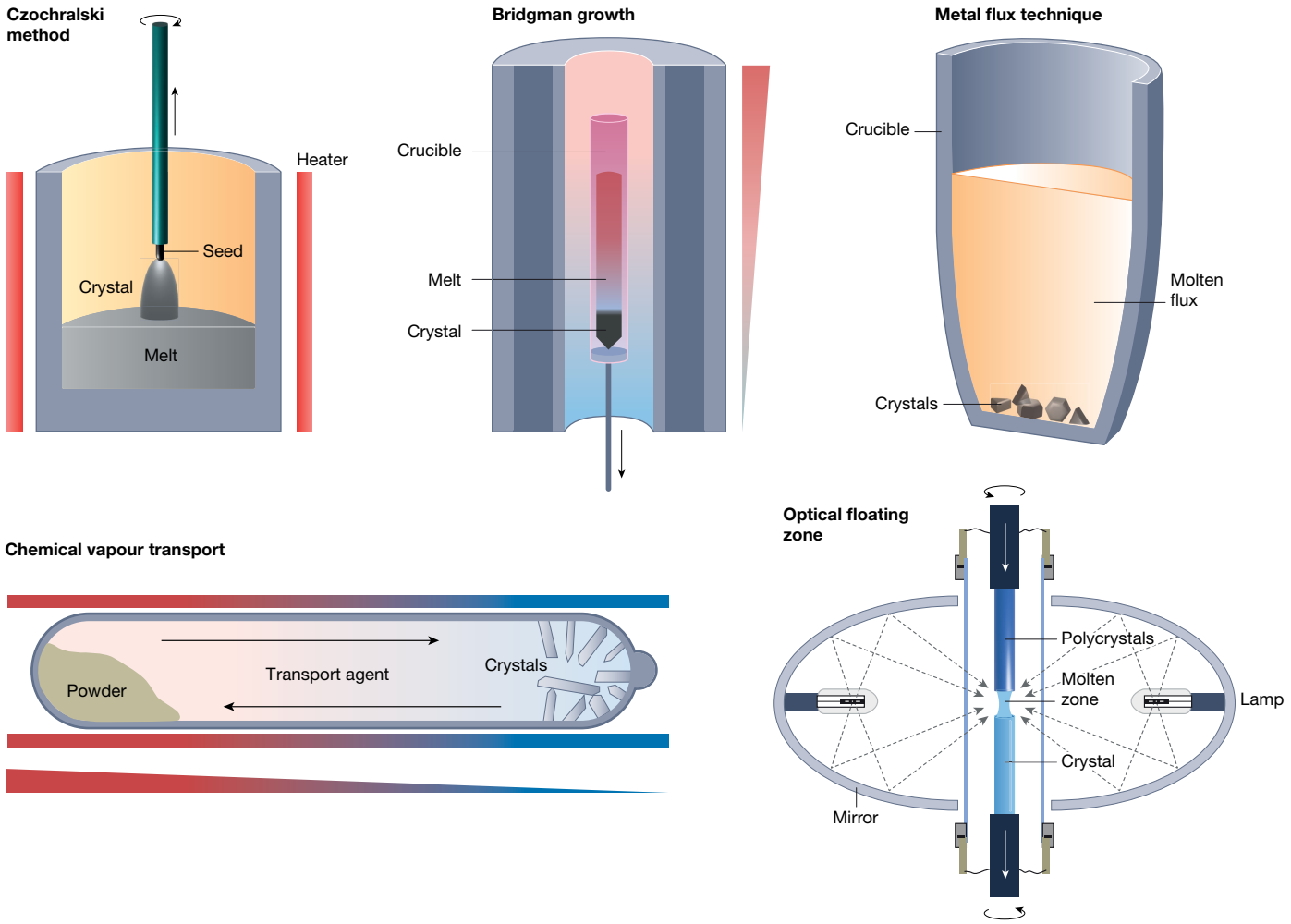

Fig. 3 | Examples of single-crystal growth techniques, reflecting a small part of the variety of the synthesis processes. Simplified sketches of the five most common crystal growth methods from the melt (Czochralski, Bridgman, optical floating zone and metal flux growth) and the gas phase (chemical vapour transport).

Once a structured database on synthesis has been established, this will allow computer-aided development of synthesis recipes to fabricate as yet unknown materials with tailored properties. Moreover, it will enable comparison of different synthesis methods for the same material in terms of generalized physical and chemical parameters, also linking them to theoretical predictions.

FAIR data in experimental disciplines

Experimental materials science is concerned with the characterization of the atomic and electronic structures of compounds, as well as with determining their electrical, optical, magnetic, thermal or mechanical properties. Typically, terabytes (sometimes petabytes) of data from one study result in a few plots in a publication. Only FAIR data management of all results, both successful and failed, makes experimental studies reproducible and obviates the necessity to repeat the experiments for a different but related project. In addition, by making all these data available to the community, everyone will benefit from statistically more reliable quantification of measurement errors and calibrations.

In experimental materials science, the variety of characterization methods is very diverse, and each class of methods has its own equipment and workflows for generating data. The diversity in data formats specific to vendors, labs, instruments, communities and operators presents a substantial challenge with regard to integrating this information into a FAIR infrastructure. For the initial period, we concentrate on five experimental techniques (see Fig. 4) with very different frontiers in terms of the 4V challenges, and largely disjunct and differently structured communities. These are electron microscopy and spectroscopy, angle-resolved photoemission spectroscopy, core-level photoemission spectroscopy, optical spectroscopy and atom-probe tomography. The amount of generated data ranges from a few kilobytes to terabytes per dataset, and the data rates and data structures also differ substantially. With some modern detectors delivering several gigabytes of data per second, the volume and velocity challenge is to preprocess, compress and evaluate or visualize these data. This becomes a more severe velocity issue in time-resolved experiments, for which the duration may not even be fixed but being decided during the observation. Disturbingly, overall, we observe a lack of efficient and reliabe recording of metadata in a digital form, posing a severe data-veracity challenge.

Analogously to establishing FAIR data in synthesis, a strong focus is the customization of inter-operating ELNs and LIMSs, their integration into experimental workflows and their direct connection to the data repository.

In each of the five selected experimental techniques, activities have also started to define domain-specific metadata catalogues and ontologies. In some labs (transmission electron microscopy and spectroscopy), a first, rudimentary prototype of a NOMAD Oasis has recently been installed, with the aim of exploring how it should be further developed towards the requirements of the different subdomains. Integration of ELNs into the experimental workflow is at different

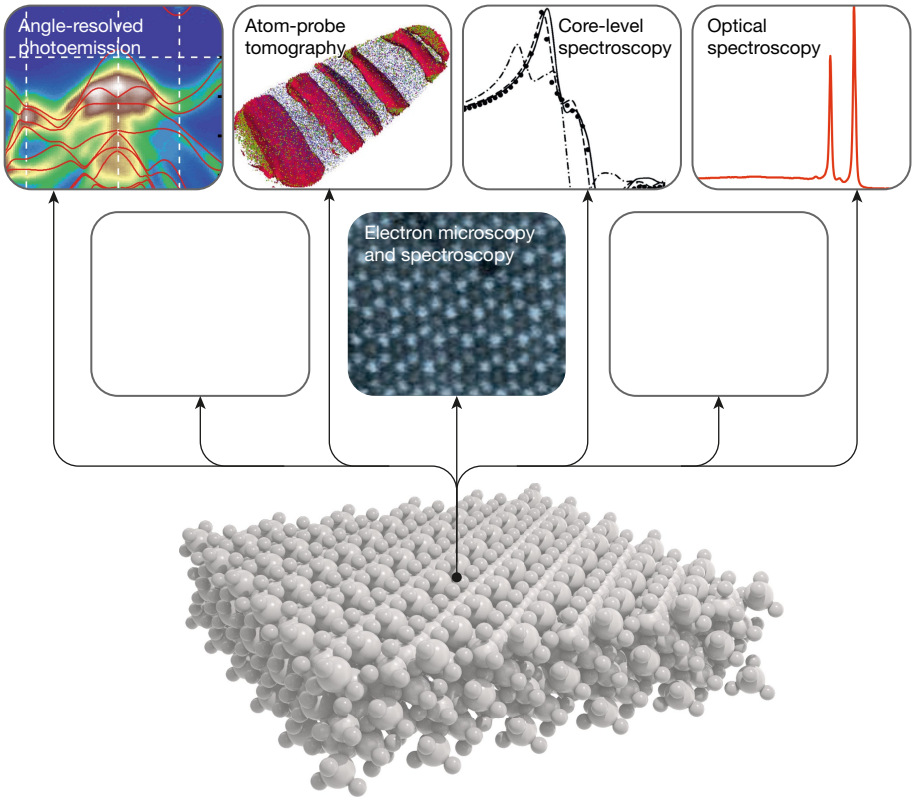

Fig. 4 | Illustration of five experimental material characterization techniques chosen to focus on initially, for the purpose of establishing digital FAIR data-management workflows. Optical spectroscopy, atom-probe tomography, angle-resolved photoelectron spectroscopy, electron microscopy and X-ray photoemission spectroscopy are noted explicitly, and the large number of other experimental methods is indicated by the empty 'perspectives'.

stages of development, ranging from first implementation concepts in atom-probe tomography to a working integration of an ELN database with the data acquisition software in some transmission electron microscopy labs. This also includes the tagging of samples with QR labels and automated linking of sample IDs with links to experimental data and time-stamped notes generated by the data acquisition software. Several angle-resolved photoemission spectroscopy groups are reorganizing their labs at present, switching from paper lab books to ELNs. In this context, we note that, in a joint effort of different labs, we were able to make vendors of complex equipment reconsider their previously restrictive and closed data-format policies.

FAIR theory and computations

Materials modelling, in particular, including digital twins, is enjoying ever-growing attention thanks to a timely combination of hardware and algorithmic developments[24]. The NOMAD Laboratory[6] has already implemented a materials data infrastructure for quantum-mechanical ground-state calculations and ab initio molecular dynamics (see the summary in the section on 'Data-centric materials science'). However, materials modelling also requires force fields and particle-based methods, to capture larger length scales and longer timescales (see Fig. 5). The implementation of such multiscale materials data infrastructure faces several outstanding challenges[25,26]. By considering trajectories, we need to account for both instantaneous and ensemble properties. Also, the heterogeneity of simulation setups, solvers, force fields and observables requires an ambitious and coherent strategy to make multiscale modelling FAIR. The development of metadata for this field has only just started.

Another crucial task is the response of matter to external stimuli. The physical objects of interest obtained from theory are excitation energies and lifetimes, electronic band gaps, dielectric tensors, various excitation spectra and ionization potentials, all of which have experimental counterparts. The leading methodologies[27] comprise time-dependent density functional theory, Green function techniques and dynamical mean-field theory, implemented in a huge number of different computer codes. The predicted FAIRmat infrastructure will foster the often incomplete documentation at present and facilitate benchmarking and curation of results.

Concerning the four Vs, the area of theory and computation is severely affected by variety (that is, the heterogeneity of the meaning of the produced data). This refers to the fact that there are many physical equations, even more algorithms and yet more approximations that are implemented in the numerous, very different software packages. Although ab initio computational materials science has largely assumed a common nomenclature, for example, for the several hundred exchange-correlation approximations (see, for example, Libxc, a library of exchange-correlation and kinetic energy functionals for density functional theory; https://tddft.org/programs/libxc/), this is not yet the case for force fields, dynamical mean-field theory or calculations of fluid dynamics, and so on.

Related to the variety challenge is veracity. Note that we differentiate between accuracy and precision; the latter can be checked by comparing results from different software that addresses the same equations and uses the same approximations. Although for ab initio computational materials science the first important steps have been made[28], for other theoretical approaches, such efforts are still missing. Accuracy, in turn, refers to the equations and basic approximations (for example, the exchange-correlation functional used or the force field). Here, error bars are largely missing so far, but for interoperability with experimental results, such error estimates need to be developed. Concerning the data volume, for molecular dynamics calculations, it is

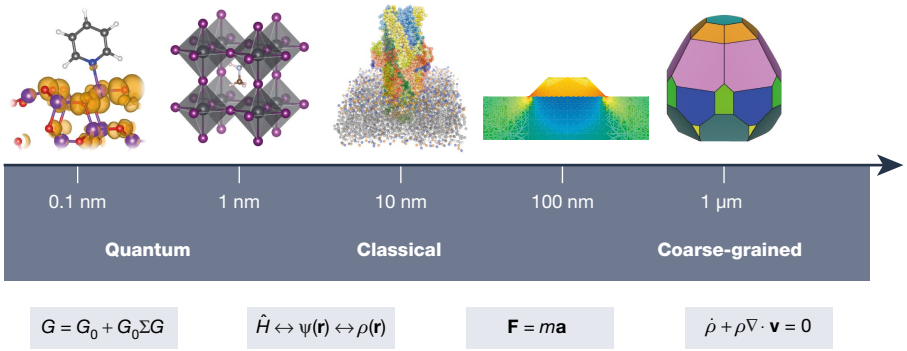

Fig. 5 | Multiscale challenge of materials simulations, ranging from sub-ångström up to length scales of micrometres and more. Approaches range from excited states, ground-state density functional theory and force-field-based molecular dynamics to continuum. The data infrastructure has to meet the needs of the different methods. These are outlined by the equations at the bottom.

hardly possible to store all the information—that is, the detailed time evolution of the positions of all the atoms (if these are several thousand in number, as in force field studies) or the electronic charge density (in ab initio studies). Here, selection and compression strategies will be developed.

Making the data revolution happen

Fourteen years after Jim Gray's explication on data-intensive scientific discovery[29], materials science is still dominated by the first three research paradigms: experiment, theory and numerical simulations[2]. However, there is now wide consensus that data-centric research and the fourth paradigm (data mining, new ways of analysis (largely by AI) and visualization) will change, if not revolutionize, the sciences. We stress that the fourth paradigm represents a new way of thinking[2]. It complements but does not replace the previous concepts and approaches. Implementation of this paradigm not only creates new opportunities but also enhances the traditional approaches through efficient data exchange, better documentation and a more detailed understanding of what other groups are doing. This will open new horizons for research in the basic and engineering sciences, reaching out to industry and society.

So, let us summarize what we need to make the data revolution happen in materials science:

- Hardware for data storage and handling, advanced analytics and high-speed networks. The availability of appropriate hardware is the basic prerequisite for building the described data infrastructure. We also need middleware, for example, for the efficient exchange of data that are created in or by different digital environments. In addition, efficient, near-real-time data analytics will also require advanced hardware, as well as software and hardware co-design.
- Development and support of software tools. New tools are already being invented; for example, for fitting data, removing noise from data, learning rules that are behind patterns in data and identifying 'statistically exceptional' data groups[16]. With such rules, one will also identify 'materials genes'—physical parameters that are related to the processes that trigger, facilitate or hinder a certain materials property or function. FAIRmat will foster the international coordination of the development of such tools in the wider materials science community.
- The development of ELNs and LIMSs. Such necessary changes of current scientific procedures seem minor if one accepts that it is good scientific practice to document the experimental (or computational) conditions and the results in full detail, so that studies are reproducible. Thus, data collection (including the comprehensive characterization of the experimental setup) should become as automatic as possible. This sounds like an outdated request, but it has not been executed properly so far and, for data-centric science, it is essential. Unfortunately, for some—maybe many—studies, an immediate realization is not fully possible and even the first approximation requires a 'phase transition'. Owing to the complexity of the field, there is no one-size-fits-all solution.
- Close collaboration between experts from data science, IT infrastructure, software engineering and the materials science domain as equal partners. In FAIRmat, this will be realized by a centralized hub of specialists at the Physics Department of the Humboldt-Universität in Berlin.
- Changing the publication culture and advancing digital libraries. As noted above, the basic scientific requirement of reproducibility of experimental work is often lacking. This is rooted in the complexity and intricacy of materials synthesis. FAIRmat will change this situation. The concept of 'clean data'—that is, data that are comprehensively annotated—is being developed (see ref. [30] and references therein). This is much more elaborate than it sounds, and publications that 'just' present and describe such data comprehensively should be appreciated by the community as much as a standard publication in a high-impact journal.

Digital libraries have been built and advanced over the past decade, and this work continues. Although there have been ample developments in the field of life sciences, the situation in materials science is less advanced. However, it is improving (for example, at https://tib.eu/ or https://openaire.eu/) and, in this field, metadata catalogues are typically too unspecific to allow the identification of suitable datasets (for example, for AI analysis).

The German National Research Data Infrastructure project (https://nfdi.de) promotes all of the points discussed above, with the exception of the necessary hardware. Although a national effort, it is obviously part of an international activity, and FAIRmat has established respective collaborations already. We will support scientists and confirm them in their responsible handling of research data, and we will strive to educate the next generation of researchers and engineers to actively engage in order to achieve these goals in a timely manner.

The field is changing and the research community seems mostly convinced about the direction of this change, but it is still mostly in the role of an observer. If active scientists don't sign on, the infrastructure will develop without them. Then, in a few years, they will need to accept what is there, and it may—unfortunately—not fully serve their needs. The consequences of the whole endeavour may be summarized as follows: the predicted changes brought about by a FAIR data infrastructure will not replace scientists, but scientists who use such an infrastructure may replace those who don't.

1. Wilkinson, M. D. et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* **3**, 160018 (2016).
This work coined the acronym FAIR, which is now used worldwide.

2. Draxl, C. & Scheffler, M. In *Handbook of Materials Modeling* (eds Andreoni, W. & Yip, S.) 49–73 (Springer, 2020).
Addresses the fourth paradigm of materials research and highlights the related challenges.
3. Editorial. Empty rhetoric over data sharing slows science. *Nature* **546**, 327 (2017).
4. Draxl, C., Illas, F. & Scheffler, M. Open data settled in materials theory. *Nature* **548**, 523 (2017).
5. Himanen, L., Geurts, A., Foster, A. S. & Rinke, P. Data-driven materials science: status, challenges, and perspectives. *Adv. Sci.* **6**, 1900808 (2019).
6. Draxl, C. & Scheffler, M. The NOMAD laboratory: from data sharing to artificial intelligence. *J. Phys. Mater.* **2**, 036001 (2019).
7. 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**, 46 (2017).
8. Talriz, L. et al. Materials Cloud, a platform for open computational science. *Sci. Data* **7**, 299 (2020).
9. Andersen, C. W. et al. OPTIMADE, an API for exchanging materials data. *Sci. Data* **8**, 217 (2021).
10. Wittenburg, P., Lautenschlager, M., Thiemann, H., Baldauf, C. & Trilsbeek, P. FAIR practices in Europe. *Data Intell.* **2**, 257–263 (2020).
11. Tanaka, I., Rajan, K. & Wolverton, C. Data-centric science for materials innovation. *MRS Bull.* **43**, 659–663 (2018).
Describes the status of data-centric research in materials science.
12. Ghiringhelli, L. M., Vybiral, J., Levchenko, S. V., Draxl, C. & Scheffler, M. Big data of materials science: critical role of the descriptor. *Phys. Rev. Lett.* **114**, 105503 (2015).
13. Isayev, O. et al. Universal fragment descriptors for predicting properties of inorganic crystals. *Nat. Commun.* **8**, 15679 (2017).
14. Jäckle, M., Helmbrecht, K., Smits, M., Stottmeister, D. & Groß, A. Self-diffusion barriers: possible descriptors for dendrite growth in batteries? *Energy Environ. Sci.* **11**, 3400–3407 (2018).
15. Ward, L., Agrawal, A., Choudhary, A. & Wolverton, C. A general-purpose machine learning framework for predicting properties of inorganic materials. *NPJ Comput. Mater.* **2**, 16028 (2016).
16. Kulik, H. et al. Roadmap on machine learning in electronic structure. *Electron. Struct.* https://doi.org/10.1088/2516-1075/ac572f Section 1.4 (2022).
17. Kraft, O., Wenzel, W. & Wöll, C. Materials research in the information age. *Adv. Mater.* **31**, 1902591 (2019).
18. Cantor, B. Multicomponent and high entropy alloys. *Entropy* **16**, 4749 (2014).
19. Chung, Y. G. et al. Advances, updates, and analytics for the computation-ready, experimental metal–organic framework database: CoRE MOF 2019. *J. Chem. Eng. Data* **64**, 5985–5998 (2019).
20. Ahmad, M., Luo, Y., Wöll, C., Tsotsalas, M. & Schug, A. Design of metal-organic framework templated materials using high-throughput computational screening. *Molecules* **25**, 4875 (2020).
21. Haldar, R. et al. A de novo strategy for predictive crystal engineering to tune excitonic coupling. *Nat. Commun.* **10**, 2048 (2019).
22. Rosen, A. S., Notestein, J. M. & Snurr, R. Q. Structure–activity relationships that identify metal–organic framework catalysts for methane activation. *ACS Catal.* **9**, 3576–3587 (2019).
23. Kim, E. et al. Materials synthesis insights from scientific literature via text extraction and machine learning. *Chem. Mater.* **29**, 9436–9444 (2017).
24. Ceder, G. & Persson, K. The stuff of dreams. *Sci. Am.* **309**, 36–40 (2013).
25. Bereau, T. Computational compound screening of biomolecules and soft materials by molecular simulations. *Model. Simul. Mater. Sci. Eng.* **29**, 023001 (2021).
26. Jackson, N. E., Webb, M. A. & de Pablo, J. J. Recent advances in machine learning towards multiscale soft materials design. *Curr. Opin. Chem. Eng.* **23**, 106–114 (2019).
27. Martin, R. M., Reining, L. & Ceperley, D. M. *Interacting Electrons. Theory and Computational Approaches* 1st edn (Cambridge Univ. Press, 2016).
28. Lejaeghere, K. et al. Reproducibility in density-functional theory calculations of solids. *Science* **351**, aad3000 (2016).
29. Hey, T., Tansley, S. & Tolle, K. (eds) *The Fourth Paradigm: Data-Intensive Discovery* (Microsoft, 2009).
30. Trunschke, A. et al. Towards experimental handbooks in catalysis. *Top. Catal.* **63**, 1683–1699 (2020).

Acknowledgements This work received funding from the Max Planck Research Network BiGmax, the Humboldt-Universität zu Berlin and the European Union's Horizon 2020 research and innovation programme under grant agreement no. 951786, the NOMAD CoE. We thank S. Auer for his feedback on digital libraries, V. Coors for her thoughtful work on the figures and the entire FAIRmat team for shaping the concept and first steps towards its implementation. FAIRmat is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation), project 460197019.

Author contributions C.D. (chairperson of FAIRmat) and M. Scheffler (co-chair) created the general concept, led the writing and edited the whole paper, in close discussion with all the other authors. They also wrote the abstract, introduction, 'Data-centric materials science' and 'Making the data revolution happen' sections. M. Albrecht and C.F. took the lead on the section 'FAIR, reproducible synthesis' (Area A of FAIRmat); M.G. and C.T.K. on the section 'FAIR data in experimental disciplines' (Area B of FAIRmat); M. Scheffler., K.K. and T.B. on the section 'FAIR theory and computations' (Area C of FAIRmat); and H.-J.B., M. Scheidgen and W.E.N. on the section 'FAIR data infrastructure for materials science' (Area D of FAIRmat). C.D., together with C.W. and A.G. (Area E of FAIRmat) and M. Scheffler and M. Aeschlimann (Area F of FAIRmat), created the section 'Preparing the research of tomorrow'. All authors carefully read and edited the whole paper.

Competing interests The authors declare no competing interests.

Additional information
Correspondence and requests for materials should be addressed to Claudia Draxl.
Peer review information *Nature* thanks Kerstin Kleese van Dam, Brian Matthews and the other, anonymous, reviewers for their contribution to the peer review of this work.
Reprints and permissions information is available at http://www.nature.com/reprints.
Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.