Perspective

Accelerating the adoption of research data management strategies

Johanne Medina,[1,*] Abdul Wahab Ziaullah,[2] Heesoo Park,[3] Ivano E. Castelli,[4] Arif Shaon,[5] Halima Bensmail,[6] and Fedwa El-Mellouhi[2,*]

SUMMARY

The need for good research data management (RDM) practices is becoming more recognized as a critical part of research. This may be attributed to the 5V challenge in big data: volume, variety, velocity, veracity, and value. The materials science community is no exception to these challenges as it heralds its new paradigm of data-driven science, which uses artificial intelligence to accelerate materials discovery but requires massive datasets to perform effectively. Hence, there are efforts to standardize, curate, preserve, and disseminate these data in a way that is findable, accessible, interoperable, and reusable (FAIR). To understand the current state of data-driven materials science and learn about the challenges faced with RDM, we gather user stories of researchers from small- and large-scale projects. This enables us to provide relevant recommendations within the data-driven research life cycle to develop and/or procure an effective RDM system following the FAIR guiding principles.

INTRODUCTION

Data are being generated exponentially by different research groups and organizations from various fields to address growing concerns about social, health, and environmental problems. The rapid advancement in computing and storage capabilities over the last decade has also been a catalyst in this data deluge. Therefore, it is not surprising that we now have increasingly voluminous and diverse datasets available at hand. The size of such a dataset could range from hundreds of terabytes to several petabytes. Fittingly, the term used to refer to these vast datasets is big data,[1] which presents modern researchers and scientists with a unique set of challenges that need to be addressed to realize the full potential of such data.

Today, we are facing the so-called the 5V challenge in big data (Figure 1) which concerns volume (amount of data), variety (non-homogeneity of data types, meaning, and sources), velocity (the rate at which data are generated), veracity (quality and accuracy of the data), and value (what users can do with the collected data).[2] According to SeedScientific,[3] as of 2021, we create a voluminous amount of roughly 2.5 quintillion (10^{18}) bytes of data daily. These data are acquired and collected from various sources in both the industry and in academia from tech-giant companies, startups, governmental organizations, research institutions, and individual users. These can be either structured or unstructured data and can take on several forms, including raw, processed, shared, or published, implying that disparate units and data types accompany them.

PROGRESS AND POTENTIAL

Materials science has heralded its new paradigm in data-driven science following the generation of big data from high-performance computing and high-throughput experimentations. Such big data need to be standardized, curated, preserved, and disseminated in a way that is findable, accessible, interoperable, and reusable (FAIR) to make use of its full potential. The materials science community is in its premature stage concerning adapting research data management (RDM) practices. In this work, we provide detailed recommendations to be followed within the data-driven research life cycle, which aims to promote RDM within the community. More interoperable materials databases and standards need to be developed and adopted within the community to get the maximum benefit from this initiative. The nature of heterogeneous data in materials science makes this a huge challenge. However, if we all, as a community, work together to make our data FAIR, materials discovery could indeed be accelerated.

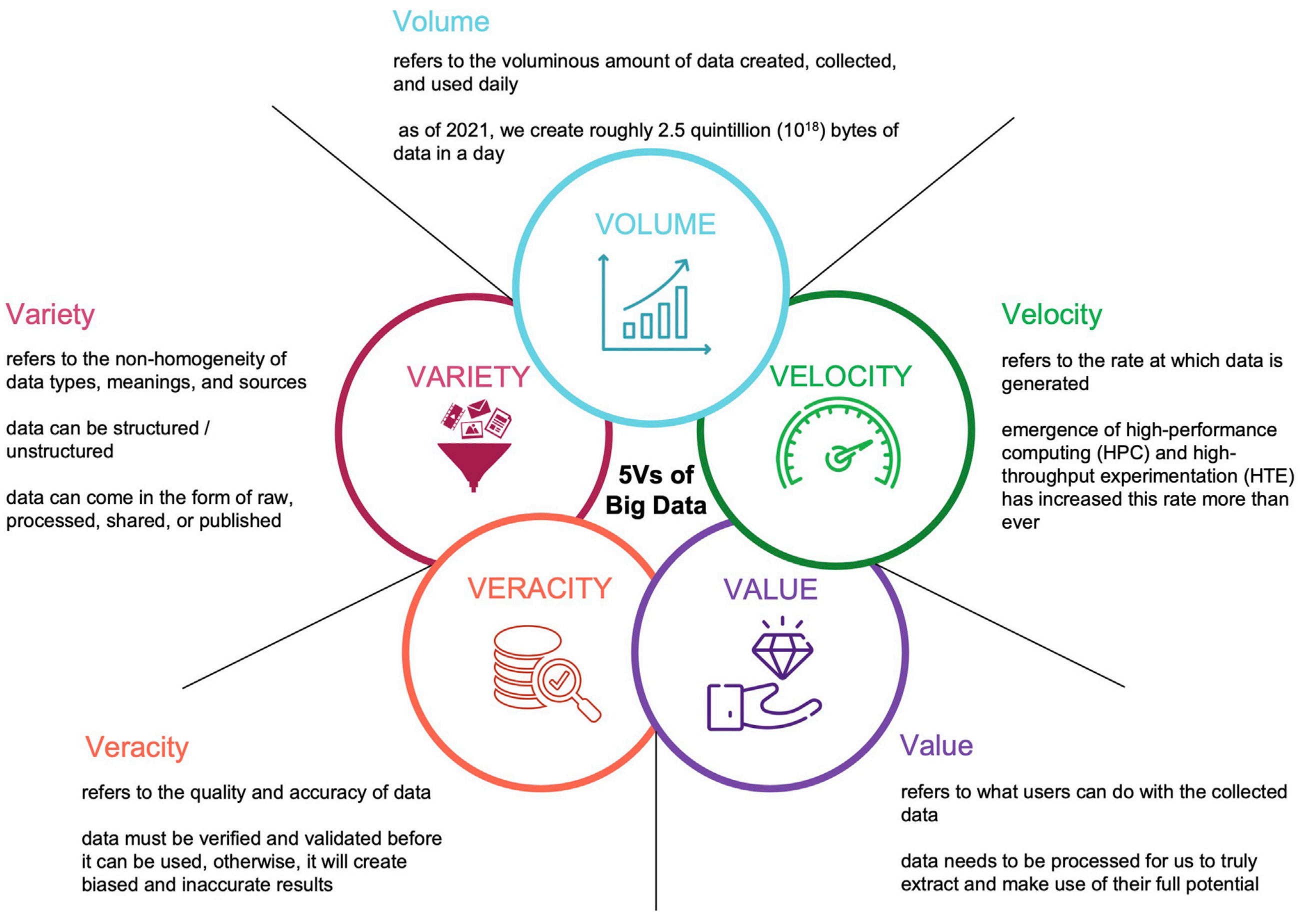

Volume
refers to the voluminous amount of data created, collected, and used daily
as of 2021, we create roughly 2.5 quintillion (10^{18}) bytes of data in a day
VOLUME
Variety
refers to the non-homogeneity of data types, meanings, and sources
data can be structured / unstructured
data can come in the form of raw, processed, shared, or published
VARIETY
5Vs of Big Data
VELOCITY
Velocity
refers to the rate at which data is generated
emergence of high-performance computing (HPC) and high-throughput experimentation (HTE) has increased this rate more than ever
VERACITY
VALUE
Veracity
refers to the quality and accuracy of data
data must be verified and validated before it can be used, otherwise, it will create biased and inaccurate results
Value
refers to what users can do with the collected data
data needs to be processed for us to truly extract and make use of their full potential

Figure 1. The 5V challenge of Big Data

Technological advancement, with the emergence and adoption of high-performance computing[4,5] and high-throughput experimentation (HTE),[6] has paved the way for data to be produced at a higher rate than ever before.

However, all these points pose a question on the quality and accuracy of the data being produced. Keeping track of and validating every data point is a tedious but necessary job that needs to be acknowledged. Data must be verified and validated before they can be used; otherwise, it will create biased and inaccurate results with huge variance.[7]

Before we can fully make sense of the problems that arise from this data explosion, it would be useful to first understand what data are and why they are becoming an increasingly important commodity. Data are considered as the new oil of the digital era, as popularly publicized by the media.[8] Oil is the driving force of the machines and industrial products we enjoy today. Similarly, data are deemed to enable the equivalent with information and communication technology by being the driving force of software applications and digital marketing now, and even more so in the future. It is one of the assets of today, if not the most important. One of the most prominent use cases of big data is in machine learning. Big data are an enormous collection of data that reveal patterns, correlations, and dependencies using machine learning, which cannot be extracted from small datasets.[9] They enable machine learning algorithms to perform better by making more accurate predictions[10]

[1]College of Science and Engineering, Hamad Bin Khalifa University, P.O. Box 34110, Doha, Qatar

[2]Qatar Environment and Energy Research Institute, Hamad Bin Khalifa University, P.O. Box 34110, Doha, Qatar

[3]Centre for Material Science and Nanotechnology, Department of Chemistry, University of Oslo, Oslo, Norway

[4]Department of Energy Conversion and Storage, Technical University of Denmark, 2800 Kgs. Lyngby, Denmark

[5]Qatar National Library, Qatar Foundation, P.O. Box 5825, Doha, Qatar

[6]Qatar Computing Research Institute, Hamad Bin Khalifa University, P.O. Box 34110, Doha, Qatar

*Correspondence: jmedina@hbku.edu.qa (J.M.), felmellouhi@hbku.edu.qa (F.E.-M.)

https://doi.org/10.1016/j.matt.2022.10.007

and reducing issues related to over-fitting[11] and sampling bias. However, such datasets need to be processed for us to truly extract and make use of their full potential. When data are processed, analyzed, and utilized efficiently, they can produce digital solutions of much greater value. Well-processed data are used in feeding machine learning models that escalate the process of pattern discovery and prediction. Many factors affect the performance of a machine learning model, but the quality and instance of the data are at the top of the list.[12] Redundant, noisy, and irrelevant data undermine the predictions made by machine learning algorithms. Therefore, data pre-processing is considered the most important and time-consuming task in this process and should not be overlooked.[13–17] Pre-processing includes data cleaning, normalization, feature extraction, feature selection, feature engineering, transformation, outlier detection, and the detection of missing and inconsistent data. These steps are essential and will significantly affect a machine learning model's efficiency and accuracy; thus, organized raw data are vital.

To mitigate these challenges, the wider research communities are increasingly recognizing the need for research data management (RDM) practices that will enable higher research impact, data reuse, and preservation for the long term.[18] This constitutes the planning, organization, curation, discovery, accessing, sharing, and publishing of research data.[19] Efficient RDM practice addresses all aspects of research datasets that are sourced, produced, and used within the life cycle of a research project, which includes, but is not limited to, records, literature, experiments, parameters, measurements, surveys, interviews, simulations, code, algorithms, software applications and their versions, packages, intermediate data results, and final output.[20] To maximize the use of these data, it is necessary to standardize, curate, preserve, and disseminate them with proper data management practices. A term that will heavily be referred to in this paper is metadata. Metadata is often understood as data about data. It answers the questions of what, when, where, how, and why data are created and utilized. The proper standardization of metadata is one of the building blocks in achieving an effective RDM.

The materials science community is no exception to all the aforementioned challenges when dealing with big data and the management of such. Currently, we are in the fourth paradigm of materials science where materials discovery is data driven, which combines experimental automation, artificial intelligence, and data management tools to revolutionize scientific research.[9,21,22] In the regime of data-driven materials science research, having well-organized and curated data is vital to leverage substantial datasets to reach a critical mass.

This white paper aims to understand the current state and challenges associated with data-driven materials science and big data management. To do so, we survey researchers working on a small-scale research project and another within a large consortium. We analyze the current research status of each community and the challenges they face regarding the use and management of research data. This enables us to provide relevant recommendations to develop and/or procure an effective RDM system following the findable, accessible, interoperable, and reusable (FAIR) guiding principles,[23] in accordance with the data-driven research life cycle.

DATA-DRIVEN MATERIALS SCIENCE RESEARCH

The world constantly needs new materials and solutions to sustain the key sectors of our economy, yet the discovery of such materials has more or less followed a

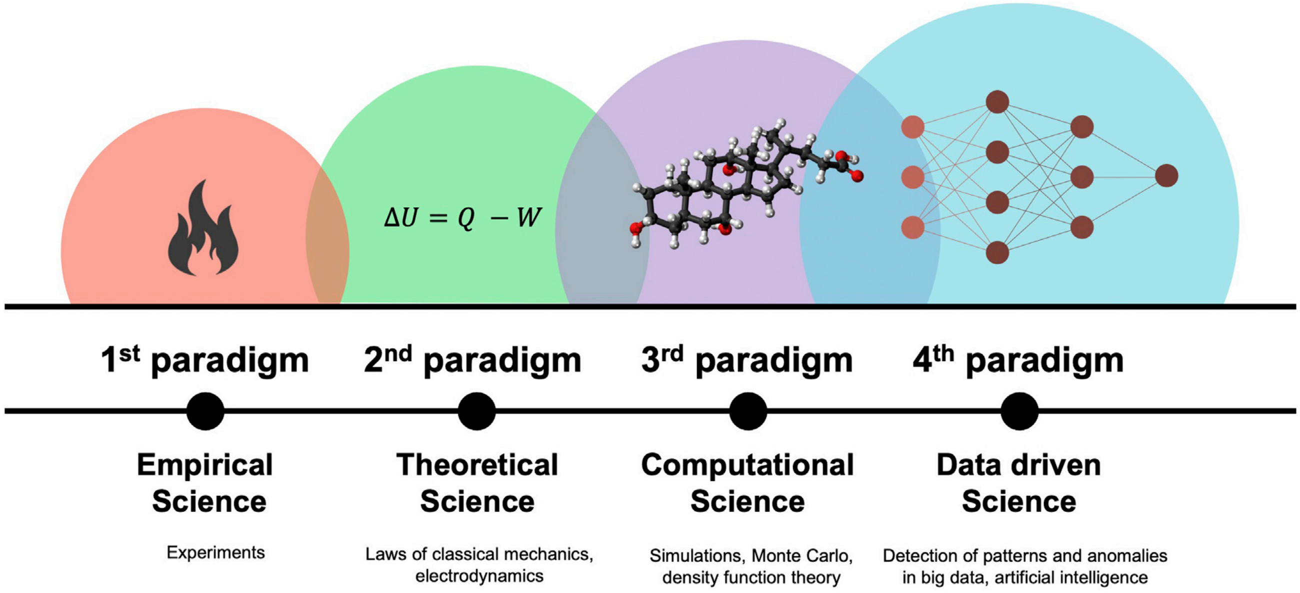

ΔU = Q − W
1st paradigm
2nd paradigm
3rd paradigm
4th paradigm
Empirical Science
Experiments
Theoretical Science
Laws of classical mechanics, electrodynamics
Computational Science
Simulations, Monte Carlo, density function theory
Data driven Science
Detection of patterns and anomalies in big data, artificial intelligence

Figure 2. The four paradigms of materials science

trial-and-error process.[22] Up until the computational science era, the average molecule-to-market lead time for new materials was 10–20 years, which is in no doubt a huge amount of time.[24] Now, materials science is entering a new regime beyond experimental and computational processes. Data science, machine learning, and artificial intelligence are revolutionizing people's everyday lives, with all communities adapting them to their work. Materials science is no exception to this trend, as it heralds its new paradigm of data-driven science[21,25,26] (Figure 2). This new paradigm represents a new way of thinking that does not replace but rather complements the previous fundamental computational and experimental approaches.[27]

The Materials Genome Initiative (MGI)[28] was inaugurated in 2011 with the aim of achieving a reduction of greater than 50% percent lead time and have materials ready for the market within 2–3 years. A huge catalyst to achieving this mission is the utilization of artificial intelligence and the adoption of data-driven science. Data-driven materials science extracts knowledge from materials datasets to automate and accelerate materials discovery and validation. Because of the relative scarcity of many types of materials data, and the inherent impracticality of gathering massive annotated datasets, a brute force strategy of collecting massive datasets, such as those used for image recognition, natural language processing, and neural translation, is untenable in materials science.[29] The possible configurations of chemical and molecular structures to achieve a new material are practically infinite.[9] Therefore, data-driven science plays a role in finding these relevant configurations through machine learning techniques in a much faster and more efficient way than trial and error. By applying data science to materials science research, we can accelerate materials discovery at a significantly faster pace.[7]

ACADEMIA EXPERIENCES

Data are used, accumulated, and created within a research project regardless of the field of study. Documents, scientific papers, simulations, and calculations are the most typical and essential artifacts produced in research.

Working with interdisciplinary teams within a research project is a common and encouraged practice aimed at solving higher-scale problems. Common projects, for example, operate as a result of collaborations of different research groups from different institutions across the globe. Data sharing is traditionally done through physical hard drives, shared institutional network, or cloud drives. However, things can get lost, forgotten, or corrupted in this practice of data sharing, aside from device-related issues such as storage damage and limit. Thus, we are faced with a need for an efficient data sharing and reuse infrastructure to allow consistent and secure data exchange between team members. Moreover, tracking of research data can be discontinued by a researcher's departure until the replacement conducts the work. There are common cases where the team has to re-do some of their work, which is inefficient and expensive, because the new research conductors should deal with untraceable collections for themselves.

In this section, we learn about the first-hand experiences of researchers concerning the challenges they face in dealing with research data and the approaches they propose to handling those challenges.

Small-scale collaboration user story: The AIPAM project

Scientific research often includes several collaborators. Even when the work of an individual collaborator may be compartmentalized or abstracted from other members, the work may still need to be aggregated into a contiguous body. Such a contiguous body can be represented by a data-flow pipeline by which the data of experiment and computation flow in different forms. In particular, as an interdisciplinary project demands an effective engagement of each researcher to other fields, the data pipelines are critical for the transferability between the fields of each expertise. Therefore, data management in the generation, transformation, and assimilation becomes a crucial part of the research as their research expands in both volume and scope.

There have emerged materials acceleration platforms such as BIG-MAP, Materials Project, and AiiDA. These platforms play a role in centralizing data flow in materials discover, while data collection and dissemination are their primary objectives.[30,31] In contrast, in small-scale collaborations, the main goal would instead be publications of their discoveries, such as newly designed materials and conceptualized chemical reactions. The task of data collection and dissemination is often prioritized at lower ranks than the main objectives. Therefore, such task re-prioritization results in obstacles when the project scales up.

For example, in our research project, Artificial Intelligence Platform for Accelerating Materials Discovery (AIPAM), we faced difficulties in inducing a new member to the team to replace/reproduce the work of an existing team member who had left the team. This project implements a data-driven high-throughput pipeline to discover optimal compositions of halide perovskites and corresponding chemical and electronic properties for photovoltaic applications. We devised the pipeline by using a plethora of Python modules and libraries, while conducting a series of research aiming at the discovery of optimal compositions of halide perovskites and corresponding chemical and electronic properties for photovoltaic applications.[32–34] This complexity and peculiarity of the project called for specific expertise and domain knowledge at different stages in the pipeline during the development period. In particular, the high-throughput density functional theory (DFT) and feature engineering stages of the pipeline required deeper insight into two different specialized domains for pre-processing and post-processing tasks. For illustration, the pre-processing step of the high-throughput DFT required chemical intuition because we were interested in the chemical composition of perovskites and their impact on the physical properties. At the same time, we needed to post-process the results to extract the key data and understand the sensitivity of computed results on varying materials as well as the theoretical methods. Building a well-organized database at this stage was critical to have a consistent dataset to train a machine learning model and to practice FAIR data curation. Subsequently, while we combined the post-processing in the high-throughput DFT and pre-processing in the feature engineering steps, finding the appropriate feature inclusion was necessary to infer the physical properties. In particular, we learned that, when we predicted the descriptor values that were estimated by using the DFT results, the workflow was more efficient and robust than when we directly predicted the DFT-computed physical values depending on the machine learning algorithms. Accordingly, the in-depth understanding of both quantum chemistry and machine learning algorithms was a prerequisite for our workflow development. Therefore, designing and implementing a back-end machine learning engine was challenging at the small-scale collaboration size because only a handful of colleagues shared such domain overlap. This crucial aspect impeded the project's successor from being familiar with the complete data generation pipeline.

One of the reasons the team struggled with formally representing the complex computational workflow[35] was that it encompassed two different domains; namely, material science and machine learning. Moreover, as the research progressed, more refinements were added to the workflow, which posed additional requirements to represent an addition or a change in knowledge.

Moreover, an additional hurdle for member's transition was attributed to the fact that the research artifacts such as (paper, code, data) needed to be consolidated by the new member, without having any systematic approach. Obtaining insights through the data could be time consuming and may require a steep learning curve due to the varying scope of the applied fields. Moreover, the papers produced alone did not contain all the information required to reproduce the results from the existing team member. The page limits in most scientific journals resulted in the paucity of the content required to sufficiently reproduce the results. This restraint in practice often occurred in reproducing the previous results for any new members.

Non-reproducibility of results is a statistically significant issue in the research community. According to a survey presented by Baker,[36] around 70% of researchers have been unsuccessful in reproducing the results from another research. In the case of AIPAM, several factors contributed to the non-reproducibility of the results, which included code revisions used, approximation criteria, and post-processing of the data. Such complexity demands more rigorous representation of knowledge to extend FAIR protocols to workflows.

All these problems faced by the AIPAM team from preserving and disseminating the intellectual property created during the research warrant a more robust system of artifacts consolidation and knowledge representation. One of the ways in which this can be done is to utilize a knowledge representation infrastructure based on ontology and metadata. Moreover, adhering to a consistent template for the input and output data mitigates any confusion and cognitive overload[37] for new members. Such a template can appropriately be defined as a part of a data management plan (DMP).

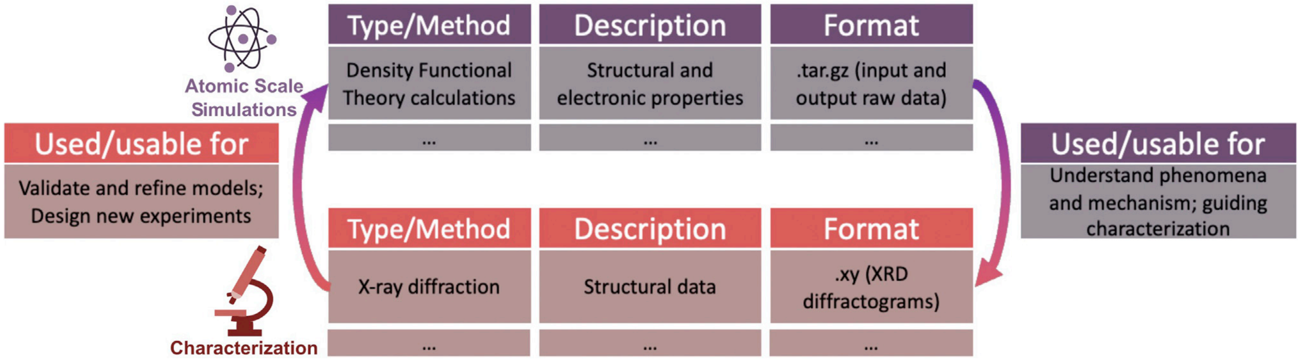

Atomic Scale Simulations
Type/Method
Description
Format
Density Functional Theory calculations
Structural and electronic properties
.tar.gz (input and output raw data)
...
...
...
Used/usable for
Validate and refine models; Design new experiments
Used/usable for
Understand phenomena and mechanism; guiding characterization
Type/Method
Description
Format
X-ray diffraction
Structural data
.xy (XRD diffractograms)
...
...
...
Characterization

Figure 3. Data tables describing the generation and flow of research data across the BIG-MAP project

Large consortia user stories: The BIG-MAP project

The development of new electrochemical storage devices is necessary to accelerate the transition to a sustainable future. Several limitations are currently hindering this transition, from the scarcity of raw materials to lack of a fundamental understanding of the phenomena happening in a battery, in particular at its interface. Numerous large-scale national and international consortia have been established to tackle the ambitious goal of accelerating battery discovery. These consortia include Battery500 in the Unites States,[38] BATTERY 2030+ in Europe,[39] the Faraday Institution in UK,[40] and Post Lithium Storage Cluster of Excellence (POLIS) in Germany.[41] These projects have in common that they establish materials acceleration platforms, which combine data from different battery domains, from multiscale models and material synthesis to characterization and cell testing. Data, intended as any research output, e.g., results, models, samples, scripts, standard operating procedures, and metadata, are used at different levels, from benchmark theoretical and experimental results to train machine learning models to predict materials properties, bridge scales, and identify underlying laws of nature. FAIR data is thus a fundamental requirement to enable this research approach. Standardization of RDM is one of the main focuses of the BATTERY 2030+ and of one of its flagship project, Battery Interface Genome – Materials Acceleration Platform (BIG-MAP).[42] BIG-MAP aims at accelerating the battery discovery through the development of a unique infrastructure and accelerated methodologies to bridge data and competences across the entire battery value chain. The core competences within BIG-MAP rely on artificial intelligence, high-performance computing, autonomous synthesis robotics, and high-throughput characterization. A detailed and comprehensive DMP is fundamental not only to have an overview of the terabytes (TB) of data produced by each of the 34 partners but, more importantly, to follow the flow of data between partners and battery domains. Figure 3 shows an example of two entries for atomistic simulations and characterization tasks and their connections.[31] Focused on what is produced, rather than how, the DMP is centered on data tables describing which data are generated and what is their type and format. Starting from these tables, is it possible to generate links between the battery chain; for example, indicating that the data generated by simulations are collected to benchmark characterization results, or characterization data are delivered to simulations to provide atomistic insight into a phenomenon, as indicated in Figure 3. Figure 4 shows the overall link between the different domains involved in BIG-MAP. To ensure their FAIRness, data are published open-source in the Materials Cloud Archive (tag BIG-MAP).[43,44] The scrips are, instead, collected as apps in the BIG-MAP App Store.[45] It is important to point out that some data (in particular, from industrial partners) are protected by intellectual property rights (IPRs). The BIG-MAP infrastructure has thus different levels of

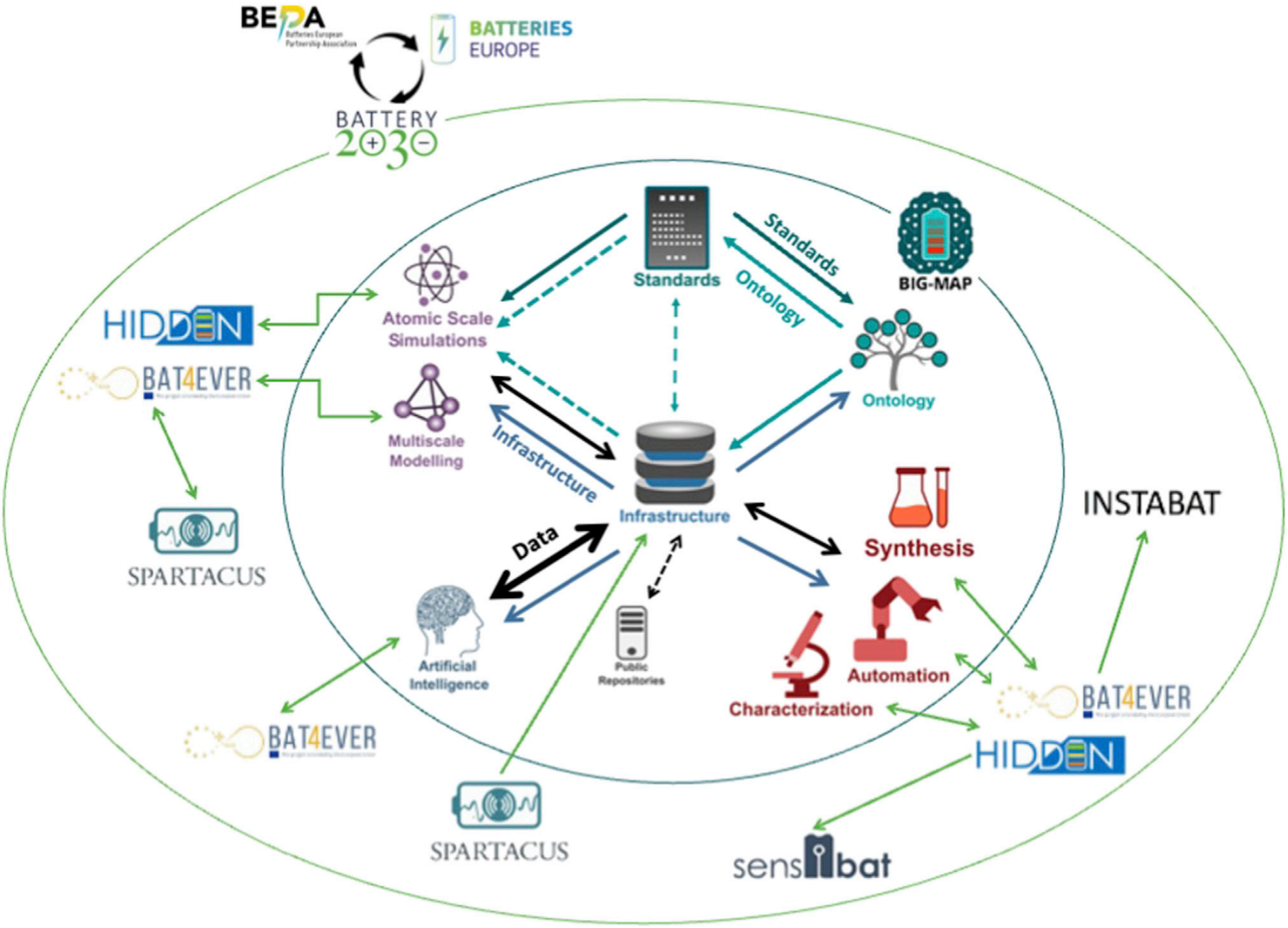


Figure 4. Flow of data between the different domains involved in the BIG-MAP project (inner circle) and the BATTERY 2030+ consortium (outer circle)
The ambition is to connect the entire battery research in Europe and beyond.

openness. The large majority of the data are open-source; however, when necessary, some data are restricted and accessible only to the consortium.

The definition of a DMP has been a group work, which involved experts from each battery domain. This allows all the techniques involved in the project to be covered as well as the data wishes from the different partners. Starting from the data produced in each task, we have been able to create a data network within the work package, project, and beyond. Because of the variety of data sources, a platform to store data from the entire battery domain is not ready yet (BIG-MAP is working in that direction). In particular, experimental and theoretical data are stored in very different ways, which are often not machine readable and interoperable. Our intermediate approach is to create an interface to allow theoretical models and autonomous experiments to talk to each other.

Beyond the BIG-MAP project, the entire BATTERY 2030+ consortium (six research projects in total) has adopted a similar DMP template. This has also allowed data to be linked across projects; e.g., the data infrastructure established in BIG-MAP can be used in the SPARTACUS project.[46] The process of identifying links is, at this point, manual, inefficient, and very tedious because it is based on visual inspection of several thousands of spreadsheet lines. By embedding ontological concepts into data, i.e. by establishing an ontology for data based on the battery ontology (BattINFO),[47,48] the consortia aim at automating the linkage between data, thus contributing to connecting all of the battery research in Europe.

Funding agencies are extremely supportive of the implementation of RDM plans. This, in fact, does not only increase re-usability of data beyond the funded project but also increase the trustworthiness of the data themselves, allowing other researchers to reproduce results or to standardize protocols and infrastructures. The European Commission, for example, established a pilot project under the Horizon 2020 program, called Open Research Data Pilot (ORD pilot),[49] aiming at increasing the awareness on RDM and their practical applicability. Similar approaches have also been taken by national funding agencies, publishers, and universities.

The community is now at a stage where defining DMPs is a requirement for most new projects, and researchers are starting to see the importance of sharing their data. However, the practical implementation of DMPs is still at its early stages. Very often, research data are stored in private repositories, lacking the basic FAIR principles, thus not fulfilling the requirements listed in the DMP. Moreover, the lack of a common data ontology, and the large number of different approaches adopted and repositories implemented, make the link between data and projects difficult. Thus, to be effective, DMPs require the efforts of the entire community to define standards and ontologies, which need to be chemistry and technology neutral. Dedicated manpower is needed for this. The largest research initiatives should take the lead in this, organize workshops, and draft memoranda of understanding, educating the new generation of scientists in how to handle data in the correct way. Research is not only final results but also the entire path to produce them.

FAIR GUIDING PRINCIPLES

To achieve efficient RDM that solves the challenges of handling big data, it is important to understand and follow the FAIR guiding principles. The FAIR guiding principles[23] aim to improve the current infrastructures used in making research data reusable. These principles describe distinct considerations for data publishing that can be understood by both humans and machines to support deposition, exploration, sharing, and reuse. They are domain-independent, high-level principles that can be applied in various research fields.[50] These principles define characteristics that data should possess in order to facilitate discovery and reuse by other members of the community. Following these principles is a vital step toward making data human and machine readable.

Findable

Making data findable is motivated by open science research. Open science research ensures that data are transparent and available to all of academia, industry, and governmental stakeholders.[51] With the amount of research projects being conducted worldwide, it is most likely that several projects are utilizing the same datasets even for different purposes or, if not explicitly the same, a close variation of one. Having said that, it would be useful and practical to be able to search for these data in a findable way in order to avoid doubling of work to save resources and energy.

Accessible

The A in FAIR means accessible. Once data are found, users must have clear information on how to access them and their respective metadata. The scientific field is slowly diverting toward open access of data, which implies that data can be accessed by anyone, not just the ones working on the project.[7,9] However, since we do not always desire that due to privacy, security, ethical, and intellectual property issues, part of data management is defining access and authorization rules to one's data. Access conditions may vary, including access on request, ethics approval may apply, public access, restrictions may apply, or temporary restriction, depending on the needs of the project or the policies of the funding entity. It is then important to note that FAIR data are not equivalent to open-access data.[52] While open access is for everyone to have all published data at their disposal, FAIR defines clear and concise rules on who has the specific authority for this privilege.

Interoperable

Interoperability is the ability of different systems, environments, and non-cooperating resources to communicate, work together, and understand each other with minimal effort.[23] As the name suggests, data should operate across, inter, or among, different applications. Data should be comparable in that they can be understood from various implementations. The same data column may be named differently but mean the same thing and serve the same purpose. Likewise, a data column may be named the same but actually mean opposite things. Machines and humans alike need to notice this distinction and make reasonable decisions based on this observation to maintain data integrity. It is important for data to be delivered with standard formats and units, much like speaking the same universal language, to maximize their benefit for users.

Reusable

The R in FAIR is defined in several ways but all essentially mean the same thing. The official definition is reusable,[23] although others also interchange it with the terms reproducible[53] and repurposable. This implies that one should be able to utilize data created from one research project or scientific question into another different application. Some materials and data can be used for different scenarios; therefore, it would save a lot of time and resources if one could use the same existing data from a known and reputable source.

Having data that are reproducible also helps in research validation and verification.[50] What is obvious to you, with the skills and knowledge you currently have, does not apply to everyone else or even to the future you within a few months or years.

RECOMMENDATIONS

Now that we understand what FAIR means and how they contribute to the success of a research project, a question that could arise is how can a research project implement the underlying principles in an operational context? The general research project life cycle for data-driven materials science is illustrated in Figure 5. Although this flow might not exactly agree with the practice of each and every data-driven material scientist, we developed this life cycle based on personal experiences. We expect that other researchers' flows will at least have a similar variation to this.

In this section, we traverse through this life cycle and briefly explain what we mean by them. We then propose a set of recommendations and best practices to guide researchers what FAIR steps to take at each phase of their research flow. By following these practices, researchers are expected to achieve proper data management and dissemination guided by the FAIR principles. These recommendations were formulated by combining the most relevant and urgent propositions from literature[50,52,54–56] and by suggesting our own based on the user stories in the previous

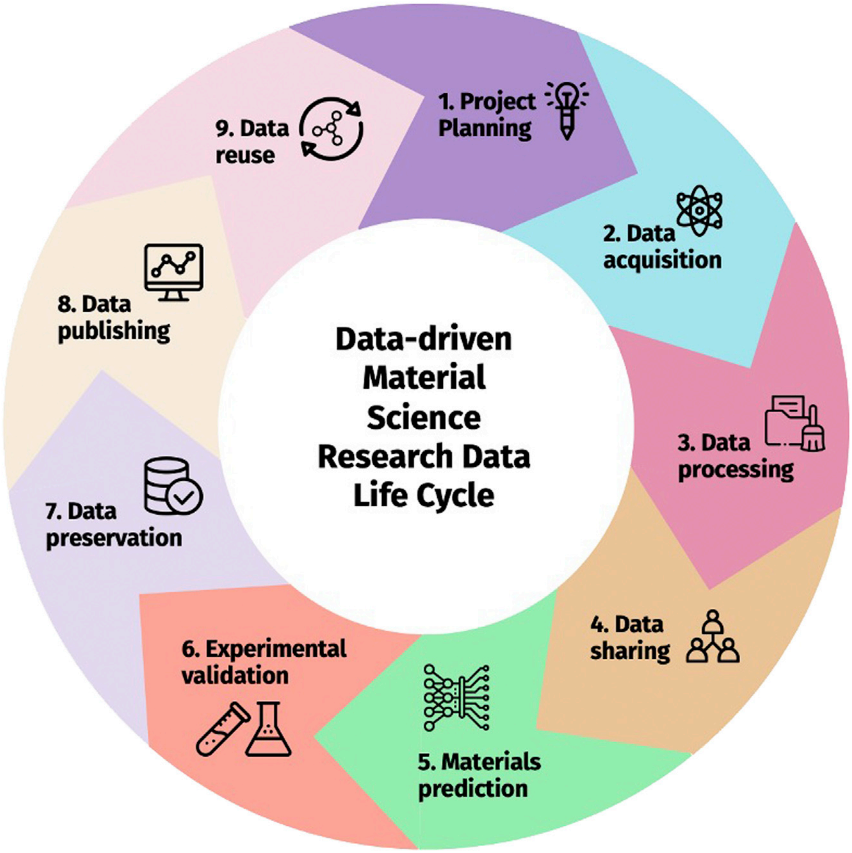


Figure 5. Data-driven materials science research data life cycle

We traverse through this data life cycle and provide recommendations for each phase on what key steps to take to achieve FAIR data.

section. The aim of these recommendations is to enable effective data sharing and reuse to preserve data value for long-term use.

Project planning

Project planning is the initial phase of research. This includes recognizing a problem and formulating a hypothesis to prove. Concurrently, responsible researchers should make several key decisions and considerations about data management as part of the planning before even starting the data creation/collection process.

Decide on the data you collect and how to store them

Data are a crucial component in the development of a research project; therefore, it is necessary to choose their properties carefully. The first step in ensuring a well-organized project is deciding on the input data's source, type, and format. In materials science, computational and experimental scientists' results, parameters, simulations, and experiments provide a valuable data source. Likewise, existing data and third-party data sources from materials databases are also accepted and encouraged. It is essential to consider the implications (including intellectual property [IP], copyrights, and ethical and legal considerations) of using these data sources, including the coverage (temporal and/or geographical) on how the data will be used and shared during and after the completion of the project. An expectation on data volume to be used within a project shall be made clear in the beginning, as this will later facilitate the structure and storage of the needed database. Last, staying consistent with these decisions and following them throughout the research life cycle is equally important.

Create and follow a detailed DMP

Efficient planning from the early stages of a research project life cycle has been proved to be the key to successful RDM that yields good-quality research data. Therefore, one of the most important activities aimed at raising awareness about research data is the creation of a detailed DMP.[31]

A DMP is a document that answers questions about how data will be managed, curated, and preserved, among other aspects of the data life cycle. A good DMP records what data are created, sourced, and analyzed and how such will be used and processed. It also includes how they will be stored and accessed and who has the authority to do such. Basically, it is a living document articulating what (type of data), why (rationale), and how (methodology) data flow throughout the entire project life cycle. In addition, aligning the DMP with previously published DMPs allows us to potentially retrieve public data and share the new data with a larger community. To do so, it is important to adhere to RDM guidelines and align with published standards and ontologies.

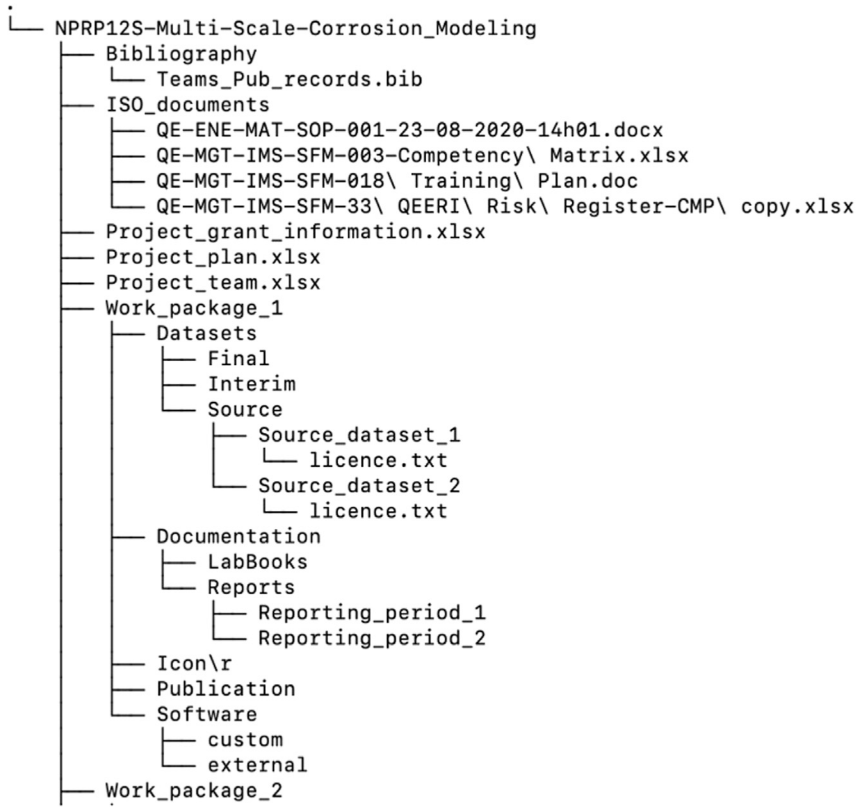
The information captured in the DMP could serve multiple purposes to ensure long-term FAIRness of a dataset, including the following:

- Contributing to the metadata capture process for both interim and final data results. Many institutions, such as the University of New South Wales, Australia,[57] provide data management tools that enable exporting metadata from plans to support data publishing and reuse.
- Providing a "live" checklist for important RDM activities throughout the research project.

Although the definition of a DMP does sound intimidating, it is really nothing more than a documentation of how you envision your data to be managed and how you plan to make it FAIR. It is relatively lightweight, with usually no more than 10 pages, and to be seen by only a limited number of people. A helpful tool to get you started is DMPOnline,[58] which can help you draft, review, and share DMPs that meet organizational and institutional requirements. Currently, an institutional requirement for DMPs around the world is still considerably rare; however, we are advocating for everyone to pick up this practice to progress into a better scientific community.

Develop a file and folder naming standard

Best practices for file and folder naming should be adopted to ensure that data are easy to locate and use. It is recommended to use meaningful names for files and folders; e.g., *projectName_softwareUsed_importantVariables_06012022.csv* is significantly better and informative than *test1.csv*. Moreover, ensure that file and folder names are machine readable with an interoperability with parsers on any computer platform. This implies that the use of special characters such as*:&$![]*{}/should be avoided as they are often used for specific tasks in a digital environment. Some of the most popular naming conventions used in coding,[59] which can also be adapted to concatenate useful information within file and folder names include *Snakecase* (e.g., project_name_date), *Pascalcase* (e.g., ProjectNameDate), and *Camelcase* (e.g., projectNameDate). Figure 6 shows the file and folder structure developed and adopted for the AIPAM project described in section "academia experiences."



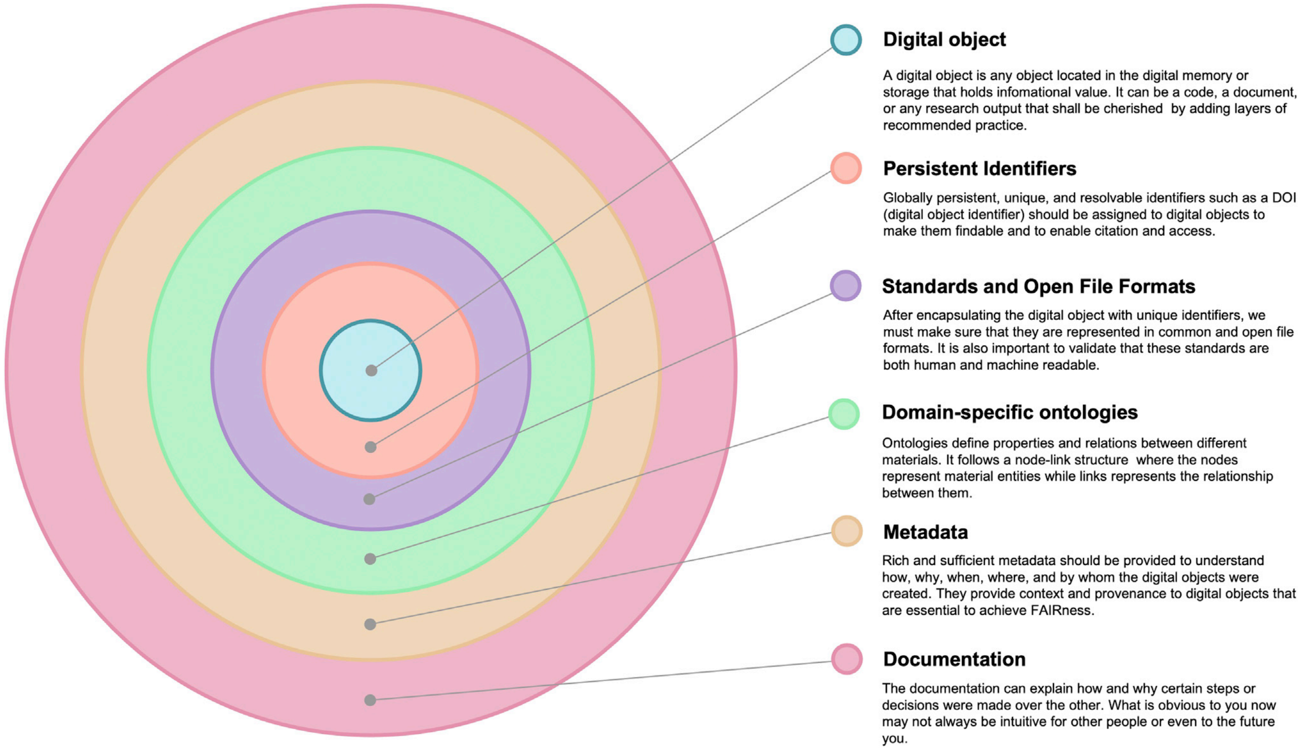
```
.
└── NPRP12S-Multi-Scale-Corrosion_Modeling
    ├── Bibliography
    │   └── Teams_Pub_records.bib
    ├── ISO_documents
    │   ├── QE-ENE-MAT-SOP-001-23-08-2020-14h01.docx
    │   ├── QE-MGT-IMS-SFM-003-Competency\ Matrix.xlsx
    │   ├── QE-MGT-IMS-SFM-018\ Training\ Plan.doc
    │   └── QE-MGT-IMS-SFM-33\ QEERI\ Risk\ Register-CMP\ copy.xlsx
    ├── Project_grant_information.xlsx
    ├── Project_plan.xlsx
    ├── Project_team.xlsx
    ├── Work_package_1
    │   ├── Datasets
    │   │   ├── Final
    │   │   ├── Interim
    │   │   └── Source
    │   │       ├── Source_dataset_1
    │   │       │   └── licence.txt
    │   │       └── Source_dataset_2
    │   │           └── licence.txt
    │   ├── Documentation
    │   │   ├── LabBooks
    │   │   └── Reports
    │   │       ├── Reporting_period_1
    │   │       └── Reporting_period_2
    │   ├── Icon\r
    │   ├── Publication
    │   └── Software
    │       ├── custom
    │       └── external
    ├── Work_package_2
```

Figure 6. File and folder naming convention developed and adopted for the AIPAM project

Data acquisition and data processing

The data acquisition phase in materials science research includes data creation, extraction, and procurement from computations and experiments. If the objective of the research is very niche, such as optimization of organic-cation- based halide perovskite,[34] the procurement of data from a material science database might not be possible due to availability issue. Moreover, obtaining data from computations and experiments can be very expensive as they require significant computation and operational complexities. This motivates adoption of active learning or data-driven machine learning to resourcefully acquire new data. Data-driven science is often utilized to tackle sparse data problems. The objective of data-driven machine learning is to form an initial surrogate model using the available data and use this model to determine what additional data would improve the accuracy of the model and minimize uncertainties. In case initial sparse data are not available, random data acquisition methods are often utilized[60,61] to form a sparse yet sufficient initial model. Such a principled approach enables the acquisition of more informative data rather than obtaining new data completely random. Several methodologies are available in the scientific literature that deal with active learning or data-driven machine learning; the two prominent ones are Bayesian optimization[62] and reinforcement learning.[63]

Once data are procured, they will undergo a phase commonly referred to as data processing. This process includes data cleaning, transformation, scaling, dimension reduction, and outlier detection. These processed data are crucial in training machine learning models for materials discovery and thus need to be efficiently represented to enable access and discovery for long-term use.



Digital object

A digital object is any object located in the digital memory or storage that holds infomational value. It can be a code, a document, or any research output that shall be cherished by adding layers of recommended practice.

Persistent Identifiers

Globally persistent, unique, and resolvable identifiers such as a DOI (digital object identifier) should be assigned to digital objects to make them findable and to enable citation and access.

Standards and Open File Formats

After encapsulating the digital object with unique identifiers, we must make sure that they are represented in common and open file formats. It is also important to validate that these standards are both human and machine readable.

Domain-specific ontologies

Ontologies define properties and relations between different materials. It follows a node-link structure where the nodes represent material entities while links represents the relationship between them.

Metadata

Rich and sufficient metadata should be provided to understand how, why, when, where, and by whom the digital objects were created. They provide context and provenance to digital objects that are essential to achieve FAIRness.

Documentation

The documentation can explain how and why certain steps or decisions were made over the other. What is obvious to you now may not always be intuitive for other people or even to the future you.

Figure 7. FAIR digital object

Define FAIR digital objects

The concept of FAIR digital objects was first introduced in 2018 from the Turning FAIR to Reality report of the European Commission's 2nd High-Level Expert Group on the European Open Science Cloud (EOSC).[52] A digital object is any object located in a digital memory or storage that holds informational value. This can be a document, a dataset, or an image and essentially includes all the research data defined above that are stored digitally. A FAIR digital object is any informational unit presented according to the FAIR principles. The report by the European Commission[52] defined a FAIR digital object as a digital object encapsulated by identifiers, standards, and metadata. In our paper, we are building on this model to make it more complete and robust. Our new representation of a FAIR digital object is illustrated in Figure 7, adding layers of domain-specific ontologies and documentation.

Create globally persistent identifiers

Globally persistent, unique, and resolvable identifiers (IDs) such as a digital object identifier (DOI) should be assigned to digital objects to make them findable and to enable citation and access. These unique identifiers serve as the identifier of the digital objects that will be used to represent and locate them. Within a database, this can be implemented by creating searchable indices for every entry that can easily be referenced.

Within a Structured Query Language (SQL) database, this implies having a primary key column for all tables, and, in MongoDB, this means representing every document with either an ObjectID or a self-defined identifier key. For more examples, a scientific publication, which is one of the digital artifacts of a project, is issued with a unique DOI once published, and this is what helps other researchers to locate and cite this work. Other common identifiers for papers include International Standard Book Number (ISBN) and International Standard Serial Number (ISSN). Inherently, authors of said publications are also identified with an Open Researcher and Contributor ID (ORCID).[64] Registering for an ORCID account is highly encouraged in the research community to be able to uniquely distinguish researchers from one another and to connect their respective contributions to their name.

Use widely adopted data standards and associated file formats to ensure interoperability

To enable seamless creation and utilization of FAIR data, data standards are essential. Data standards are guidelines by which data are described and recorded, including documented agreements on representation, format, definition, structuring, manipulation, use, and management of data.[65] At present, such standards in materials science are still premature[27] but are continuously being developed, as we will further understand in the following recommendations on ontologies and metadata. The use of open file formats is highly recommended for long-term use, preservation, and interoperability as these files are usually maintained by a standards organization and can be used by anyone. Table 1 summarizes the common file formats used in all fields of research, including materials science.

For materials science molecule structure modeling, the most common and highly supported file formats include the chemical table file format family (.mol, .sd, .sdf), Chemical Markup Language (.cml), XYZ (.xyz), Crystallographic Information File (.cif), and Protein Data Bank (.pdb) format. These file formats are highly supported in different simulation software and materials repositories. It is recommended preserve your output files in one or more of these formats not only for long-term use but also for interoperability as these are some of the most widely used formats in the field. Indeed, it is impractical to save them in all literal formats, therefore the options depend on the goals of the project. For example, if the goal is for visualization using JMOL[67] or VESTA,[68] CIF and XYZ file formats are readable and supported but POSCAR is not. Moreover, Gaussian outputs (.gif) can only be read with Gaussview,[69] which limits their access and interoperability. In these cases, it is then preferred to save them in another format or convert them to another more open format (elaborated in the recommendation in section "experimental validation").

Support and develop domain-specific ontologies

One of the potential problems that can arise in collaborative research is adhering to formal naming conventions,[70] system of measurements,[71] and means to formally represent research knowledge. The former issue of naming conventions and system of measurements can be circumvented by team-wide implementation of standard nomenclature/taxonomy[72] and a system of measurements. However, representing research knowledge is still a subject of active research. Unlike taxonomy, where mere formal categorization of an entity is established, knowledge representation requires a more elaborate schema. It often demands additional structure, such as relation between various entities, rules for combining or interacting with entities, and associating various properties. Ontology provides a mechanism to establish such rigorous representation of knowledge.[73]

In computer science, ontology-based knowledge representation pertains to representing digital objects, relation between those objects, and inference rules governing interactions.[74] Domain-specific ontologies, such as materials ontology,[7] organizational workflow ontology,[75] gene ontology,[76] and battery ontology[47,48] can be adopted to

Table 1. Closed versus open file formats for text documents, spreadsheets, images, videos, and presentations[66]

File type	Instead of using (closed)		Consider using (open)	
	Name	File extension	Name	File extension
Text documents	MS Word	.doc	open document format	.odf
			HTML	.html, .htm
			plain text	.txt
			PDF	.pdf
Spreadsheets	Excel, Numbers	.xls, .xlw, .numbers	open document spreadsheet	.odf
			comma separated	.csv
Images	Photoshop, HEIC Image	.psd, .HEIC	JPEG, PNG	.jpg, .png
			GIMP XCF	.xcf
			TIFF	.tiff
			scalable vector graphics	.svg
			bitmap	.bmp
			GIF	.gif
Videos	Windows Media Video	.wmv	MPEG1, MPEG4	.mpeg, .mp4, mpeg4
	Quicktime	.mov	DivX, Ogg Theora, Dirac	.divx, .ogv
Presentations	Powerpoint, Keynote	.ppt, .pps, .key	PDF	.pdf
			HTML	.html, .htm
			open document presentation	.odp
Databases	MS Access Database File, Oracle Trace Map File	.mdb, .trm	database file	.DB
			JSON	.json
			eXtensible Markup Language	.xml
			comma separated	.csv

We recommend the use of open file formats for long-term use, preservation, and interoperability. PDF, portable document format; TIFF, tagged image file format; JSON, Javascript object notation.

relate and interlink domain-specific objects to another. This can help in automating the linkage between data across several research projects in the same domain and form a precursor to metadata. In the materials science domain, ontology-based database and metadata enable enhanced search for material properties by enabling semantic queries.[77] This can potentially lead to shortening of the time required for a researcher to look for a material with desired properties.

A recent work on ontological consideration to scientific workflows is presented by Celebi et al.,[78] with the objective of enhancing reproducibility of results in scientific research in general. Although the research on scientific-workflow-related ontology is currently premature, our case-based survey suggests that adopting standard domain-specific ontology such as BattINFO in the BIG-MAP project or developing specific case-based ontology that accommodates peculiarities of one's research can encourage better collaboration and help mitigate cognitive burden related to reproducing/validating the results of one's peers in both large- and small-scale collaborations.[79]

Define rich metadata

Rich and sufficient metadata should be provided to understand how, why, when, where, and by whom the digital objects were created. They provide context and provenance to digital objects that are essential to achieve FAIRness on all aspects.[80] Unlike data formats, which may be only machine readable, metadata are designed to be human readable and, by implementing a schema, also machine readable.

Much like ontology, it is recommended to establish standard taxonomy and schema in metadata to encourage interoperability across several contributors of the research. According to American Library Association's Committee on Cataloging: Description & Access, "A metadata schema provides a formal structure designed to identify the knowledge structure of a given discipline and to link that structure to the information of the discipline through the creation of an information system that will assist the identification, discovery, and use of information within that discipline"[81]

A key example of this can be found in materials science, where metadata standards such as the Crystallographic Information Framework (CIF)[82] provide a metadata-based schema to represent information pertaining to crystal structures. This includes standardization for representing space groups, unit cell parameters, and measurement conditions such as temperature at which the reading was taken. Another instance of standard metadata schema is the IPTC (International Press Telecommunications Council) Photo Metadata Standard,[83] which enables adding information depicted in the images, such as about people, places, and items, using a framework consisting of fields, descriptions for how fields should be used, and relevant information to be included. Such a framework enables researchers to ensure that sufficient information of their data is included.

The need for metadata schema is also substantiated in sections "small-scale collaboration user story: The AIPAM project" and "large consortia user stories: The BIG-MAP project" where the collaborators could validate the work of other team members by independently reproducing the results. However, in scientific collaboration, the data generation process often mutates as the data generation workflow is continually improved or optimized. This mutable aspect of the data generation pipeline is reflected in the data being generated and poses challenges in standardizing the schema of metadata. An example of a mutable data generation pipeline in the AIPAM project is presented in Figure 8, where blocks such as data transformation and approximation source can be altered, added, or removed depending upon the progress of the research. These dynamic changes, if accounted in the metadata, can aid researchers to utilize[84] or reproduce[53] and validate the data with ease.

There are several studies that aimed to formalize the metadata for various domains,[85],[86] and a common trait among them is utilization and formalization of taxonomy to represent key objects. For more complex domains, such as scientific workflows or digital objects across the World Wide Web, ontology-based graphs could be used. In efforts to establish semantic web by representing meta-information on digital objects in the World Wide Web, the Resource Descriptive Framework (RDF) has been established. In RDF, meta-information on digital objects, or resources, is materialized with predicates, which pertains to some aspect of the resource and properties that provide some associated value. For example, a resource such as RDM.pdf can be represented with certain predicates such as file size and a property of that predicate such as 1,000 kb. A collection of RDF statements with predicates and properties of any given resource is synonymous with building an ontology-based graph representing associated knowledge on that resource. Although it might be quite difficult for researchers to develop and maintain ontology-based metadata for small-scale collaborations, for large-scale collaborations, the benefit of utilizing ontology-based metadata outweighs the complexity associated with training researchers and maintaining it.

Write detailed documentation

To highlight the importance of documentation in the FAIRification process, we decided to add this as a separate layer to the FAIR digital object (Figure 7).

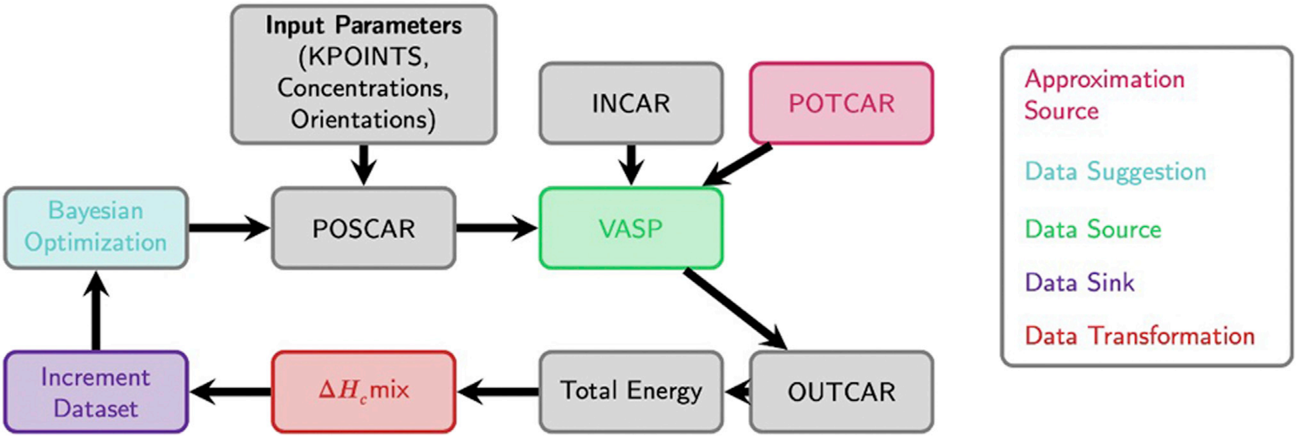


Figure 8. A mutable data pipeline for AIPAM project

Documentation can explain how and why certain steps or decisions were made over others. What is obvious to you now may not always be intuitive for other people or even to the future you. Therefore, documenting these decisions would be of great help in data re-usability and validation. This can include adding relevant comments on your code or maintaining an activity log for your daily activities as a researcher. Moreover, researchers should ensure that meeting minutes are recorded for every meeting conducted with the team. Not only does this ensure that every attendee is on the same page after the discussions but also this can be a good way to realize what key decisions were made over time.

Automate the generation of metadata, persistent identifiers, services, and ontologies

One of the most common reasons why most research groups are not following RDM practices is because scientists already have so much on their plates. Therefore, it would greatly help to automate the creation of the fundamental requirements of a FAIR digital object. The creation of persistent identifiers (PIDs) can be programmed within the user interface as datasets are uploaded to the secure database, for example. Moreover, metadata generation can be automated using parsers that extract information from a data standard.[87]

Preserve raw data independently

A secure and separate copy of the raw data shall be preserved at all times. This copy shall not be touched or manipulated directly for any process. There are cases when important data points may accidentally be deleted or modified in the processing stage; therefore, having a secure copy of the raw data ensures that researchers have room for such errors and may always go back and retrieve the original data at any point.

Data sharing

We have seen from the user stories in section "academia experiences" how common multiscale and interdisciplinary teams are in research. Data acquired from one work package may be needed in another that is not necessarily within the same institution or even country. Usually, data acquisition is performed by experimental or computational scientists, while machine learning modeling is established by computer scientists.[29] Thus, a data-sharing phase to facilitate this communication between the three parties is typical.

The data of concern at this phase are intermediate and active data results that are expected to undergo more processing or validation and thus are, for the most part, confidential. These are usually unpublished data critical in ensuring the novelty of a researcher's work, for example. The task of sharing such confidential data can be approached in one of two ways depending on the needs of the project and the policies imposed by their funding institutions. On one hand, researchers can opt to develop their own data-sharing platform to ensure a more specific and personalized experience. The other option is to use a third-party infrastructure designed specifically for private local data access, such as the NOMAD Oasis,[88] Figshare,[89] Materials Project Contribution Platform (MPContribs),[90] and DLHub.[91]

On-premise development of a data-sharing platform

Whether we like it or not, international relations do affect the extent of open and accessible science. Sanctions between countries, for example, may force researchers from a country to lose access to an existing infrastructure in use if this is hosted by a sanctioning country. Google Workspace, for example, has restricted access in Crimea, Cuba, the so-called Donetsk People's Republic and Luhansk People's Republic, Iran, North Korea, and Syria.[92] Therefore, being dependent on the services of a private company from another country has low risks in terms of sudden restrictions but is not impossible. Hence, it is useful to host your own infrastructure (including database as we will see in section "data storage, preservation, publishing, and reuse") within your own country or institution where you have full control on access and flexibility, to some extent. Advantages of this approach include the following:

- Freedom in personalizing the database structure, infrastructure components, and access control
- Minimum risks of sudden interruption of services

Disadvantages of this approach include the following:

- Additional workforce (computer scientists, full stack developers, data engineers) required for the development of the platform and the database

In this section, we provide recommendations on how to responsibly create your own on-premise data-sharing platform while ensuring consistently FAIR data.

Develop an efficient data-sharing platform with a user-intuitive Web portal

Building a browser-based infrastructure is one of the most common ways of making data findable. This is because this option gives the least amount of overhead for people trying to search for data.[7] This way, anyone with an internet connection can find data by putting in a website link on the Web browser, unlike desktop or mobile applications, where one would need to download and allocate some amount of storage memory to keep things running. It is essential to design this website considering the user experience. It should be user-friendly and intuitive enough to be used without external support. This can be achieved by providing a Help page or displaying instructions on each relevant component. The most common and recommended frameworks for Web development are React, Angular, and Vue, which are all based on the programming languages HTML, CSS, and Javascript.

An efficient data-sharing platform is a website containing all the needed data required to enable collaboration in a small-scale environment or within a consortium. It should allow for data searching, filtering, sorting, uploading, downloading, updating, and deleting. With an efficient data-sharing platform supported with complete metadata and documentation, the problem faced with the departure of a team member can be mitigated. All the previous data produced can be accessed from this platform in a timely manner. The search capability is expected to operate beyond chemical formulas or material labels but rather support discovery by surface layer, adsorption layer, catalysis, synthesis recipes, etc.[93]

Create an application programming interface

An application programming interface (API) allows for communication between the front end and the back end within an infrastructure. They enable automatic data crawling and automated access through programmable queries,[7] given that they are accompanied with a well-documented structure and metadata. With an API, data can be retrieved from code that can be written in any programming language without prior knowledge of how the database operates internally. It helps to decouple the data with a domain-specific client or Web portal. Example languages for building an API include Python Flask, FastAPI, ASP.net, Django, and node.js. The Open Databases Integration for Materials Design (OPTIMADE) consortium was created to design and implement API standards to enable seamless access and interoperability across materials databases.[94] Researchers or software developers can integrate OPTIMADE into their Web development process or choose to still develop their own API, but it is recommended to follow the specifications developed by OPTIMADE to ensure interoperability.

Display intelligent visualization

Visualizations help data be understood not only by the experts in the field but also by other users that are not very knowledgeable about the topic. They provide helpful insights into the data, opening doors to data repurposing. We have also continually highlighted that we are preparing data to be fed to machine learning models aimed at accelerating materials discovery. Part of pre-processing is data cleaning, by which visualizations such as plots and histograms are integral in detecting patterns, trends, and outliers. Creating visualizations from pre-processing to post-processing to publishing reveals important conclusions that may not be achieved without this step. The most common languages used in data pre-processing are Python and R. Both offer plotting and visualization libraries that help in exploratory data analysis.[95,96]

For publishing in chemistry and in materials science, a common tool used for visualizing chemical structures in the front-end is Jmol. Jmol is an open-source Java viewer for chemical structures in 3D. Its HTML5/Javascript counterpart is JSmol,[97] which can be integrated within a React, Angular, or Vue project. Another alternative for 3D chemical structure modeling used in website development is ChemDoodle.[98] Figure 9 illustrates the difference in how chemical structures are displayed using Jmol and ChemDoodle.

The integration of a JSmol molecule into a Javascript website is much more straightforward based on personal experience. Due to its long-term existence in the market, it has gained so much popularity that a plethora of documentation and tutorials on its implementation can be found online. On the other hand, ChemDoodle is a relatively new library and may seem to be more intimidating for integration. However, seeing how visually more realistic and pleasing it looks, projects such as Winther et al.[99] opt for the use of this library instead. The choice depends entirely on the preference of the researchers.

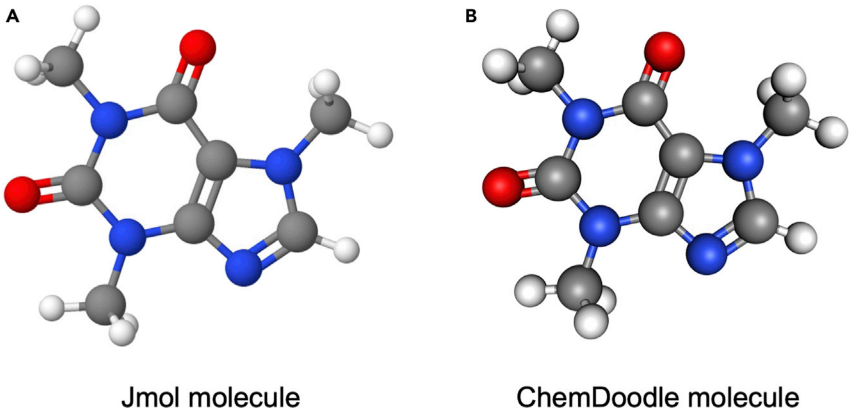


Figure 9. Visualization tools comparison

Use a third-party private data-sharing infrastructure

The NOMAD Center of Excellence (CoE) offers a private local data infrastructure called the NOMAD Oasis.[27] It can be operated as a stand-alone server with all the functionality of uploading and sharing data within the user's institutional network. It offers to hold data for faster or more private access and can specifically be used in academia to store data that a researcher does not want to publish yet. The advantages of using this approach include the following:

- Less overhead in the development and use of the data-sharing platform
- Direct use of NOMAD's established metadata

However, it poses the following disadvantages:

- Minimum control over the features of the infrastructure
- May be less specific to the needs of the project

Materials prediction

This is the phase for implementing artificial intelligence and machine learning to accelerate materials discovery. The following recommendations are specifically catered for preserving programs written to run such models.

Implement version control

Researchers should use version control systems to keep track of file changes over time while documenting when and by whom these changes were made. This allows researchers to realize how their work has progressed over time and helps new team joiners know where to start understanding things. If a team is writing code, a good repository that allows for version control is GitHub.[100] Suppose one wants to develop their own version control within the website they are developing; in that case, they can implement this by starting with saving logs of all the operations performed within a page per user. Using a linear, single-chain version control technique is recommended where one can view changes in previous links but cannot make any modifications on any version other than the current one. This means that, within a single working branch, committed versions cannot be edited anymore as this will cause version forking that is difficult to manage. Changes can only be made on the latest, most current, working version of the code within that branch. Usually, there is a master branch that is the final working directory of the code project where different branches contribute and get merged after passing a code quality check.

Capture the environment

To fully replicate an analysis later and get the same results as before, one will need to live in the same environment as that time. This means that they need the exact version of the tool or software used, the same operating system, and the same version for all the prerequisite packages and libraries of that tool or software. This is especially relevant for software development, where certain libraries get deprecated while other libraries continue to be improved and developed, causing a peer dependency error due to backward incompatibility. For this reason, we recommend working in a self-contained computing environment such as a Docker container that can be assembled anywhere.[101] Working on Docker eliminates the ''But it works on my machine!'' problem. It allows for long-term data preservation because if, in 5 years, they would want to run a pipeline within a Docker container, they would still be able to do so, given that the whole environment is captured.

Experimental validation

In order to truly verify the findings from computational and data-driven materials science, experimental studies are performed.[22,102] There is a disconnect between experimental and computational communities due to data and vocabulary siloes.[56] Therefore, facilitating a seamless comparison between the two provides a giant leap toward material discovery validation.[103] In future, materials science aims to achieve a situation where *in silico* experiments are calibrated with real-world experimental data with the least overhead possible.

Convert between various file formats

Parsers can be used to retrieve standardized metadata that aid in interoperability. With metadata parsing, we can get an artifact's complete information, allowing us to convert between file types used across different domains. Multiscale teams are in serious need of efficient parsers to enable coherent data conversion between file types. Some tools available for file format conversion are the following:

- Pymatgen[104] is a Python library that supports data parsing and conversion between file types. For example, it allows file type conversion from a Gaussian file (.gjf) used in chemistry to the coordinate system file type (.xyz) used in computational materials science.
- OpenBabel[105] is a toolkit for converting between chemical file formats. Data from molecular modeling, chemistry, solid-state materials, and biochemistry can be searched, converted, analyzed, and stored with OpenBabel. With its current version, it has support for 118 formats in total, can read 88 formats, and can write 89.
- NOMAD parser[87] converts raw data into a code-independent representation of the NOMAD Meta Info, the standard metadata structure used by NOMAD CoE.[106]

Data storage, preservation, publishing, and reuse

Digital data preservation consists of a sequence of practices aimed at ensuring that data and metadata remain FAIR on existing computer technology over a sustained period of time.[107] It is important for researchers to understand however, that data backup does not equate to data preservation.

Storing research data and the associated metadata generated and collected during a research project life cycle, securely in a durable and accessible form, is a critical requirement of efficient RDM. Responsible researchers must ensure that data are stored in a manner that meet all legal and confidentiality requirements while ensuring authenticity and integrity. One of the key decisions related to data storage is predicated on the distinction between the type of data that should be kept during the project life cycle and the data to be kept even after the project has completed. This is where the distinction between the ''active'' data (i.e., the data that are being processed or worked on during the project) and the ''final'' data (the results at the end of the project) should be understood. For example, while in some cases it may not be necessary to keep records of all incorrect or inconclusive or null results of an experiment, in other cases those types of results are as important as the final results for providing a clear account of the research undertaken. A responsible researcher should assess how the final data results along with the associated research workflow including data results may be used in other research projects before discarding any data.

Active data

For active data during the project, there may be several, both internal and external (owing to the proliferation of cloud-based storage solutions as a cost-effective way to store large volumes of data) storage solutions to choose from. The key considerations for active data are discussed next.

Have automatic backups in your database. Select a database or repository that support automatic backups. It would be safe to enable weekly or daily backups depending on how much data are uploaded in a day.

Build a scalable and secure database based on the data you have. Due to the heterogeneity of materials data and the different file formats used across different implementations (especially in experimental science[27,29,108]), there appears to be a need for more open and interoperable materials databases.[109] The existing databases in the market may not provide sufficient means for research groups to store and preserve their data; therefore, here, we provide recommendations on how people can start building their own databases to meet their needs and contribute to the community.

In general, data can be structured or unstructured, and, based on this, a database could be relational or non-relational. If data are structured with a pre-defined format, data type, and relationship for every entry in a table, then one should use a relational database. In a relational database, all rows in a table shall follow all the constraints and relationships of each corresponding column; in other words, it has a schema that every row should comply with. Some famous examples of relational databases are Microsoft SQL Server, PostgreSQL, MySQL, MariaDB, and SQLite. On the other hand, if data are unstructured such that there is no defined format or data type for each data point, then one should use a non-relational database. In a non-relational database, every document (equivalent to a row in a relational database) can have disparate columns and data types. This is often more challenging to manage. Some of the most-used examples of a non-relational database are MongoDB, Apache Cassandra, IBM CLoudant, and Couchbase.

To maintain the integrity of the data, we should have secure databases and repositories. This can be achieved by deploying said database in a secure and trusted server, either on premise or on the cloud using Web services, and by implementing rules for authorization access. While building the database, also keep in mind to consider its scalability on the expansion and continuation of a research project.[110,111]

Make confidential fields anonymous, if applicable. Allow anonymity within the database to account for confidential data that shall not be publicized following ethical and legal regulations imposed by either the government or the funding agencies. This is most commonly practiced when dealing with personal health data, but, in materials science, this might be necessary for non-published data intended only to be shared exclusively within team members.

Data anonymity can be achieved either by encrypting or hashing sensitive data using cryptography. The difference between the two is that encryption is a two-way function involving public and private keys, while hashing is a one-way function that changes plain text into a unique combination of characters. Xu et al.[112] provide a review on the different encryption standards widely used today that can be applied to our confidential data. Software developers can encrypt the input texts or files before uploading them to the database and only provide the key to decrypt to authorized users, programmed within the API layer.

Final data

For the final data result, researchers should understand the fundamental differences between data backup and digital preservation. Digital preservation is a series of continuous activities/operations (both pre-emptive and reactive) that ensure both the data and associated metadata remain FAIR on existing computer technology over a sustained period of time.[107]

Having an efficient data backup or archiving solution that typically comes with modern IT infrastructure (e.g., Microsoft Azure,[113] Amazon Web Service [AWS][114]) alone is not sufficient to mitigate the risks associated with technological obsolescence. There are many documented cases[115] from libraries and memory organizations suffering significant data loss despite having a seemingly efficient data backup and/or archival solution in place. Suitable digital preservation solutions along with a traditional backup or archival strategy is needed to ensure long-term accessibility and continued usability of research data, especially the final data results, including detailed metadata about the research workflow and, in some cases, any software applications, algorithms, or computational models developed during the project.

Upload your data on established and well-maintained repositories. Researchers may consider any of the Web-based, freely available, long-term storage solutions for data preservation and publication, including the following:

- Generalist or all purpose; e.g., Harvard Dataverse,[116] DataHub by MIT,[117] Figshare,[89] and Zenodo[118]
- Materials Science community specific; e.g., NOMAD Repository and Archive,[119] Materials Project,[120] AFLOW,[121] Catalysis-Hub,[122] JARVIS,[123] Crystallography Open Database (COD)[124] and Citrine Informatics[125]

The Registry of Research Data Repositories[126] provides a list of discipline-specific repositories.

- Institutional repository: any repository and archival solution recommended and/or provided by their institution.

These types of storage or repository solutions are considered trusted[127] and preferred to cloud-based file hosting services such as Dropbox or Google Drive, and multi-purpose software repositories, such as GitHub, as the trusted solutions offer better preservation, discoverability, and scholarly communications, such as citation tracking and impact analysis for research data. Use these platforms to publish your data as well as to retrieve data from other researchers. The aim of these infrastructures is to facilitate FAIR data handling in the easiest way possible.

Other recommendations

Consider DMPs as a part of research assessment

To encourage more researchers to follow and appreciate having proper RDM, it is the funding bodies' job to consider DMPs as part of their research assessment. Beyond the mandates of funding bodies, researchers should recognize and utilize the greater values of DMPs as an effective tool for documentation and metadata capture that may reduce the effort needed to prepare the final dataset for publishing and sharing. Currently, the major funding bodies in the United States (e.g., National Science Foundation [NSF][128]), United Kingdom (e.g., Engineering and Physical Sciences Research Council [EPSRC][129]), and Qatar (e.g., Qatar National Research Fund [QNRF][107]) require a DMP as part of any new funding applications to encourage and support implementation of good data management practices.

Contribute to data management and materials science consortia and working groups

Data management and materials science consortia are huge catalysts in improving the adaptation of FAIR data in the materials science community. GO FAIR,[130] Research Data Alliance (RDA),[131] Committee on Data (CODATA),[132] and FAIRsharing.org[133] are some of the leaders in RDM, targeting all areas of research. They aim to promote a global collaboration between researchers for efficient data sharing and reuse guided by the FAIR principles. They provide data management best practices and recommendations, active working groups, summer schools, workshops, and seminars, among other things. This makes them great resources when starting out with following RDM within your group, or even as experts in the field wanting to contribute more to the community. Examples of domain-specific consortia and platforms for materials science include the Acceleration Consortium,[134] the NOMAD Consortium,[119] the FAIRmat Consortium,[135] the OPTIMADE Consortium,[94] the USC Materials Consortium,[136] and Citrine Informatics.[137] These consortia provide many opportunities for scientists to collaborate and infrastructures to utilize.

Provide sustainable funding and reward FAIR data stewards

We can infer from all these recommendations that reaching FAIR data comes with a price. Therefore, sustainable funding shall be rewarded to teams making an effort to make this happen. Funding institutions should reward research teams who actively participate in efficient data management practices. This will encourage more teams to "jump on the train" and take these actions more seriously. The development of a single metadata standard alone, for example, was estimated to cost around $40,000 according to the leaders of the Netherlands Organization for Health Research and Development (ZonMw).[138]

Organize and attend RDM training for researchers

Encourage organizing and attending RDM training for researchers and everyone involved within a research project, including students. Skills in data science and data stewardship are required to effectively achieve FAIR data.

Include data management in the curricula for degrees

It should be within the institution's goals to include data management within the curricula for the next generation of chemists, physicists, material scientists, and computer scientists. If we invest in the young researchers of today, we may fully eliminate the problems related to data management in the future.

CONCLUSIONS

In this white paper, we outlined key practices to perform at every stage of a research life cycle to ensure that data remain FAIR and readily available for use in data-driven research. Making data FAIR requires concerted efforts from all key stakeholders, including researchers, administrators, and other contributors in a research project. However, as implied by the daily hurdle researchers face in big data management, the benefits outweigh the costs. Sufficient understanding of the underlying concepts and planning from the beginning of a research project can enable RDM activities to be incorporated within the project workflow to produce good-quality, reusable dataset.

Low-hanging fruits for effective data management could be achieved by implementing some of the relevant principles at a work package or a small project level that can be quickly and easily adopted even for projects with a modest budget. Research projects with considerable financial support within large consortia should consider the following:

- Developing and providing mechanisms and services for the storage, safekeeping, and deposition of research data in support of current and future access to research data during and after the completion of research projects
- Providing access to services and infrastructures for the storage, safekeeping and archiving of research data and records, enabling researchers to exercise their responsibilities in relation to managing their research data in line with funder policies and the responsible conduct of research

A key barrier to establishing effective good RDM practices in any project is the lack of awareness and understanding of its importance. Establishing and fostering knowledge and proficiency in RDM practices is imperative for building a sustainable research data ecosystem in any discipline. Beyond the initiatives from funding bodies, research institutions along with other key stakeholders should work together to groom champions of RDM practices through coordinated training within the curricula for the next generation of chemists, physicists, materials scientists, and computer scientists. Adhering to the FAIR principles by following the recommendations from this paper promotes collaboration and articulated communication between team members of a research group. Such practice enables acceleration in research progress, considering that data are exchanged and can be understood among all collaborators involved in less time. Suppose data are standardized and their corresponding metadata provides sufficient information to enable reuse and validation. In that case, users of such data are on the right path to accelerating materials discovery and reaching critical mass.

ACKNOWLEDGMENTS

This work is supported by the Qatar National Research Fund (QNRF) through the Graduate Sponsorship Research Award Cycle 8 (GSRA8-L-2-0503-21026) and the National Priority Research Program Cycle 12 (NPRP12S-0209-190063). F.E.-M. acknowledges support (AIPAM) from the Office of the Hamad Bin Khalifa University Vice President of Research under grant number VPR-TG01-006. I.E.C. acknowledges support from the European Union's Horizon 2020 research and innovation programme under grant agreement no. 957189 (BIG-MAP) and no. 957213 (BATTERY 2030+). H.P. acknowledges the support of MoZEES, a Norwegian Centre for

Environment-friendly Energy Research (FME), co-sponsored by the Research Council of Norway (project no. 257653).

AUTHOR CONTRIBUTIONS

Conceptualization, J.M., A.S., and F.E.-M.; investigation and resources, J.M., A.W.Z., H.P., I.E.C., and F.E.-M.; writing – original draft, J.M., A.Z., H.P., I.E.C., and A.S.; writing – editing and revisions, J.M., A.W.Z., A.S., and F.E.-M.; visualization, J.M.; supervision and funding acquisition, A.S., H.B., F.E.-M.

DECLARATION OF INTERESTS

The authors declare no competing interests.

REFERENCES

1. Sagiroglu, S., and Sinanc, D. (2013). Big data: a review. In Proceedings of the 2013 International Conference on Collaboration Technologies and Systems (CTS), pp. 42–47.

2. Naeem, M., Jamal, T., Diaz-Martinez, J., Aziz Butt, S., Montesano, N., Imran Tariq, M., De-la Hoz-Franco, E., and De-La-Hoz-Valdiris, E. (2022). Trends and future perspective challenges in big data. In Advances in Intelligent Data Analysis and Applications, J.-S. Pan, E.B. Valentina, and C.-M. Chen, eds. (Springer Singapore), pp. 309–325.

3. Vuleta, B. (2021). How Much Data Is Created Every Day? +27 Staggering Stats. https://seedscientific.com/how-much-data-is-created-every-day/.

4. T. Lynn, J.P Morrison, and D. Kenny. Heterogeneity, High Performance Computing, Self-Organization and the Cloud edited by.Springer Nature

5. Correa-Baena, J.P., Hippalgaonkar, K., van Duren, J., Jaffer, S., Chandrasekhar, V.R., Stevanovic, V., Wadia, C., Guha, S., and Buonassisi, T. (2018). Accelerating materials development via automation, machine learning, and high-performance computing. Joule *2*, 1410–1420.

6. Shevlin, M. (2017). Practical high-throughput experimentation for chemists. ACS Med. Chem. Lett. *8*, 601–607.

7. Himanen, L., Geurts, A., Foster, A.S., and Rinke, P. (2019). Data-driven materials science: status, challenges, and perspectives. Adv. Sci. *6*, 1900808.

8. Alobaidy, M. (2021). Data, the new oil of the digital era. https://www.arabnews.com/node/1825021/data-new-oil-digital-era.

9. Draxl, C., and Scheffler, M. (2020). Big Data-Driven Materials Science and its FAIR Data Infrastructure (Springer International Publishing).

10. Zhou, L., Pan, S., Wang, J., and Vasilakos, A.V. (2017). Vasilakos. Machine learning on big data: opportunities and challenges. Neurocomputing *237*, 350–361.

11. Jabbar, H., and Khan, R.Z. (2015). Methods to avoid over-fitting and under-fitting in supervised machine learning (comparative study). Computer Science, Communication and Instrumentation Devices *70*.

12. Kotsiantis, S.B., Kanellopoulos, D., and Pintelas, P.E. (2006). Data preprocessing for supervised leaning. Int. J. Comput. Sci. *1*, 111–117.

13. Wang, S., Celebi, M.E., Zhang, Y.D., Yu, X., Lu, S., Yao, X., Zhou, Q., Miguel, M.G., Tian, Y., Gorriz, J.M., and Tyukin, I. (2021). Advances in data preprocessing for bio-medical data fusion: an overview of the methods, challenges, and prospects. Inf. Fusion *76*, 376–421.

14. Alshdaifat, E., Alshdaifat, D., Alsarhan, A., Hussein, F., and El-Salhi, S.M.F.S. (2021). The effect of preprocessing techniques, applied to numeric features, on classification algorithms' performance. Data *6*, 11.

15. Graff, K., Tansey, R., Ip, A., Rohr, C., Dimond, D., Dewey, D., and Bray, S. (2022). Benchmarking common preprocessing strategies in early childhood functional connectivity and intersubject correlation fmri. Dev. Cogn. Neurosci. *54*, 101087–102022.

16. Carlos Vladimiro, G.Z. (2019). Towards explaining the effects of data preprocessing on machine learning. In 2019 IEEE 35th international conference on data engineering (ICDE) (IEEE), pp. 2086–2090.

17. Alam, S., and Yao, N. (2019). The impact of preprocessing steps on the accuracy of machine learning algorithms in sentiment analysis. Comput. Math. Organ. Theor. *25*, 319–335.

18. Banko, L., and Ludwig, A. (2020). Fast-track to research data management in experimental material science-setting the ground for research group level materials digitalization. ACS Comb. Sci. *22*, 401–409.

19. Manu, T.R. (2018). Researchers' perceptions on research data management: a survey.

20. Qatar National Research Fund. (2021). Research Data Management Plan Guidelines. https://www.qnrf.org/en-us/Funding/Policies-Rules-and-Regulations/Data-Management-Plan.

21. Agrawal, A., and Choudhary, A. (2016). Perspective: materials informatics and big data: realization of the "fourth paradigm" of science in materials science. Apl. Mater. *4*, 053208.

22. Cole, J.M. (2020). A design-to-device pipeline for data-driven materials discovery. Acc. Chem. Res. *53*, 599–610.

23. Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.W., da Silva Santos, L.B., Bourne, P.E., et al. (2016). Comment: the fair guiding principles for scientific data management and stewardship. Sci. Data *3*, 160018.

24. United States; Federal Government. (2011). Materials Genome Initiative. https://www.mgi.gov.

25. Pyzer-Knapp, E.O., Pitera, J.W., Staar, P.W.J., Takeda, S., Laino, T., Sanders, D.P., Sexton, J., Smith, J.R., and Curioni, A. (2022). Accelerating materials discovery using artificial intelligence, high performance computing and robotics. npj Comput. Mater. *8*, 84.

26. Ye, W., Zheng, H., Chen, C., and Ong, S.P. (2022). A universal machine learning model for elemental grain boundary energies. Scripta Mater. *218*, 114803.

27. Scheffler, M., Aeschlimann, M., Albrecht, M., Bereau, T., Bungartz, H.J., Felser, C., Greiner, M., Groß, A., Koch, C.T., Kremer, K., et al. (2022). Fair data enabling new horizons for materials research. Nature *604*, 635–642.

28. Materials Genome Initiative for Global Competitiveness, 2011.

29. DeCost, B.L., Hattrick-Simpers, J.R., Trautt, Z., Kusne, A.G., Campo, E., and Green, M.L. (2020). Scientific ai in materials science: a path to a sustainable and scalable paradigm. Mach. Learn, Sci. Technol. *1*, 033001.

30. Jain, A., Ong, S.P., Hautier, G., Chen, W., Richards, W.D., Dacek, S., Cholia, S., Gunter, D., Skinner, D., Ceder, G., and Persson, K.A. (2013). Commentary: the materials project: a materials genome approach to accelerating materials innovation. Apl. Mater. *1*, 011002.

31. Castelli, I.E., Arismendi-Arrieta, D.J., Bhowmik, A., Cekic-Laskovic, I., Clark, S., Dominko, R., Flores, E., Flowers, J., Ulvskov Frederiksen, K., Friis, J., et al. (2021). Data management plans: the importance of data management in the big-map project. Batter. Supercaps *4*, 1803–1812.

32. Park, H., Mall, R., Ali, A., Sanvito, S., Bensmail, H., and El-Mellouhi, F. (2020). Importance of structural deformation features in the prediction of hybrid perovskite bandgaps. Comput. Mater. Sci. *184*, 109858.

33. Park, H., ADNAN, A.L.I., Mall, R., Bensmail, H., Sanvito, S., and El Mellouhi, F. (2021). Data-driven Enhancement of Cubic Phase Stability in Mixed-Cation Perovskites (Machine Learning: Science and Technology).

34. Park, H., Kumar, S., Chawla, S., and El-Mellouhi, F. (2021). Design principles of large cation incorporation in halide perovskites. Molecules *26*, 6184.

35. Leipzig, J., Nüst, D., Hoyt, C.T., Ram, K., and Greenberg, J. (2021). The role of metadata in reproducible computational research. Patterns *2*, 100322.

36. M. Baker. 1, 500 scientists lift the lid on reproducibility. Nature, 533, 2016.

37. Gulson, R. (2021). Using schema theory to reduce cognitive load in stage 4 equation solving. Teaching Mathematics *46*, 27–33.

38. Liu, P., and Meng, S. (2022). Battery500 Consortium: Development of High Capacity Cathodes and Robust Solid Electrolytes (University of California). Technical report.

39. Amici, J., Asinari, P., Ayerbe, E., Barboux, P., Bayle-Guillemaud, P., Behm, R.J., Berecibar, M., Berg, E., Bhowmik, A., Bodoardo, S., et al. (2022). A roadmap for transforming research to invent the batteries of the future designed within the european large scale research initiative BATTERY 2030. Adv. Energy Mater. *12*, 2102785.

40. The Faraday Institution. Powering Britain's battery revolution. https://www.faraday.ac.uk/research/.

41. Polis. Post lithium storage cluster of excellence. https://www.postlithiumstorage.org/en/polis.

42. BIG-MAP. https://www.big-map.eu.

43. Talirz, L., Kumbhar, S., Passaro, E., Yakutovich, A.V., Granata, V., Gargiulo, F., Borelli, M., Uhrin, M., Huber, S.P., Zoupanos, S., et al. (2020). Materials cloud, a platform for open computational science. Sci. Data *7*, 299.

44. Materials Cloud. The BIG-MAP data repository in materials cloud. https://archive.materialscloud.org/search?page=1&size=20&q=%22BIG-MAP%22&q0=BIG-MAP.

45. BIG-MAP. The BIG-MAP app store. https://big-map.github.io/big-map-registry/.

46. SPARTACUS. https://www.spartacus-battery.eu.

47. Clark, S., Bleken, F.L., Stier, S., Flores, E., Andersen, C.W., Marcinek, M., Szczesna-Chrzan, A., Gaberscek, M., Palacin, M.R., Uhrin, M., and Friis, J. (2021). Toward a unified description of battery data. Adv. Energy Mater. *12*, 2102702.

48. Battinfo: the ontology for the battery interface genome - materials acceleration platform (big-map). https://www.big-map.eu/dissemination/battinfo.

49. European Commission. Open research data pilot in horizon 2020. https://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-data-management/data-management_en.htm.

50. Russell, K. (2021). 5 Fair Principles and Why They Matter. Digital Transformation of the Laboratory: A Practical Guide to the Connected Lab (Wiley), https://doi.org/10.1002/9783527825042.ch5.

51. Vicente-Saez, R., and Martinez-Fuentes, C. (7 2018). Open science now: a systematic literature review for an integrated definition. J. Bus. Res. *88*, 428–436.

52. European Commission (2018). Directorate-General for Research and Innovation, Turning FAIR into reality: final report and action plan from the European Commission expert group on FAIR data (Publications Office). https://data.europa.eu/doi/10.2777/1524.

53. Leipzig, J., Nüst, D., Hoyt, C.T., Ram, K., and Greenberg, J. (2021). The role of metadata in reproducible computational research. Patterns *2*, 100322.

54. Koers, H., Bangert, D., Hermans, E., van Horik, R., de Jong, M., and Mokrane, M. (2020). Recommendations for services in a fair data ecosystem. Patterns *1*, 100104.

55. Aykol, M., Hummelshøj, J.S., Anapolsky, A., Aoyagi, K., Bazant, M.Z., Bligaard, T., Braatz, R.D., Broderick, S., Cogswell, D., Dagdelen, J., et al. (2019). The materials research platform: defining the requirements from user stories. Matter *1*, 1433–1438.

56. Quay, A.N., Fiske, P.S., and Mauter, M.S. (2022). Recommendations for advancing fair and open data standards in the water treatment community. ACS ES. T. Eng. *2*, 337–346.

57. ResData. https://resdata.unsw.edu.au/resdata/.

58. Digital Curation Centre (DCC). https://dmponline.dcc.ac.uk/.

59. Research Computing University of Colorado Boulder. Coding best practices. https://curc.readthedocs.io/en/latest/programming/coding-best-practices.html.

60. Caro, G.A.D., and Abdul Wahab, Z.Y. (2021). Map learning via adaptive region-based sampling in multi-robot systems. In International Symposium Distributed Autonomous Robotic Systems (Springer), pp. 335–348.

61. Abdul, W.Z., Chawla, S., and El-Mellouhi, F. (2022). Faux-data injection optimization for accelerating computational discovery of materials.

62. Shahriari, B., Swersky, K., Wang, Z., Adams, R.P., and De Freitas, N. (2016). Taking the human out of the loop: a review of bayesian optimization. Proc. IEEE *104*, 148–175.

63. Sutton, R.S., and Barto, A.G. (2018). Reinforcement Learning: An Introduction (MIT press).

64. Haak, L.L., Fenner, M., Paglione, L., Pentz, E., and Ratner, H. (2012). Orcid: A System to Uniquely Identify Researchers*25* (Learned Publishing), pp. 259–264.

65. United States Environmental Protection Agency (EPA) (2022). EPA Data Standards. https://www.epa.gov/data-standards.

66. Akkana. Closed vs open file formats. https://shallowsky.com/openformats/.

67. Jmol. http://jmol.sourceforge.net/.

68. VESTA. https://jp-minerals.org/vesta/en/.

69. Gaussian. https://gaussian.com/gaussview6/.

70. Lowry, P.B., Curtis, A., and Lowry, M.R. (1973). Building a taxonomy and nomenclature of collaborative writing to improve interdisciplinary research and practice. J. Bus. Commun. *41*, 66–99.

71. Lloyd, R.; CNN Interactive Senior Writer (1999). Metric Mishap Caused Loss of Nasa Orbiter (CNN Interactive).

72. Duin, A.H. (1990). Terms and tools: a theory and research-based approach to collaborative writing. Bull. Assoc. Bus. Commun. *53*, 45–50.

73. Miguel-Angel Sicilia. (2013). Handbook of Metadata, Semantics and Ontologies (World Scientific).

74. Karin, K.B., Casanova, M.A., and Walter, T. (2007). Ontology in computer science. In Semantic Web: Concepts, Technologies and Applications, pp. 17–34.

75. Anzures-García, M., Sánchez-Gálvez, L.A., Hornos, M.J., and Paderewski-Rodríguez, P. (2018). A workflow ontology to support knowledge management in a group's organizational structure. Comput. Sist. *22*, 163–178.

76. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al. (2000). Gene ontology: tool for the unification of biology. Nat. Genet. *25*, 25–29.

77. Zhang, X., Hu, C., and Li, H. (2009). Semantic query on materials data based on mapping matml to an owl ontology. Data Sci. J. *8*, 1–17.

78. Celebi, R., Rebelo Moreira, J., Hassan, A.A., Ayyar, S., Ridder, L., Kuhn, T., and Dumontier, M. (2020). Towards fair protocols and workflows: the openpredict use case. PeerJ. Comput. Sci. *6*, e281.

79. Voigt, S.P., and Kalidindi, S.R. (2021). Materials graph ontology. Mater. Lett. *295*, 129836.

80. Witten, I.H., Bainbridge, D., and Nichols, D.M. (2010). Chapter 6 - metadata: elements of organization. In How to Build a Digital Library, Second Edition, I.H. Witten, D. Bainbridge, and D.M. Nichols, edsThe Morgan Kaufmann Series in Multimedia Information and Systems (Morgan Kaufmann), pp. 285–341.

81. American Library Association (2000). Task Force on Metadata: Final Report. Committee on Cataloging: Description and Access. https://www.libraries.psu.edu/tas/jca/ccda/tf-meta6.html.

82. Punla, C.S., and Farro, R.C.; Bataan Peninsula State University Dinalupihan (2022). Are we

there yet?: an analysis of the competencies of BEED graduates of BPSU-DC. Int. Multidiscip. Res. J. *4*, 50–59.

83. IPTC Standard. Photo metadata: iptc core specification version 1.1/iptc extension specification version 1.1. Doc Rev *1*.

84. Sen, A. (2004). Metadata management: past, present and future. Decis. Support Syst. *37* (1), 151–173.

85. Ashino, T. (2010). Materials ontology: an infrastructure for exchanging materials information and knowledge. Data Sci. J. *9*, 54–61.

86. Martin, T.H., Francisco Morgado, J., Goldbeck, G., Iglezakis, D., Konchakova, N.A., and Schembera, B. (2021). Domain-specific metadata standardization in materials modelling. In Domain Ontologies for Research Data Management in Industry Commons of Materials and Manufacturing.

87. NOMAD Repository and Archive (2020). Using the NOMAD parsers. https://nomad-lab.eu/prod/rae/docs/client/parsers.html.

88. NOMAD. https://www.nomad-coe.eu/about-oasis.

89. Figshare. Store, share, discover research. https://figshare.com.

90. Materials Project (2022). Introduction to MP's contribution platform MPContribs. https://docs.materialsproject.org/services/mpcontribs.

91. DLHub. https://www.dlhub.org/.

92. Google Support. https://support.google.com/a/answer/2891389.

93. Collaboration hub (2022) (Qatar Foundation Copyright).

94. Andersen, C.W., Armiento, R., Blokhin, E., Conduit, G.J., Dwaraknath, S., Evans, M.L., Fekete, Á., Gopakumar, A., Gražulis, S., Merkys, A., et al. (2021). Optimade, an api for exchanging materials data. Sci. Data *8*, 217.

95. Liu, Q., Qiao, Z., and Lv, Y. (2021). Pyvt: a python-based open-source software for visualization and graphic analysis of fluid dynamics datasets. Aero. Sci. Technol. *117*, 106961.

96. Rayan, B., and Rayan, A. (2017). Avogadro program for chemistry education: to what extent can molecular visualization and three-dimensional simulations enhance meaningful chemistry learning? World Journal of Chemical Education *5*, 136–141.

97. Hanson, R.M., Prilusky, J., Renjian, Z., Nakane, T., and Sussman, J.L. (2013). Jsmol and the next-generation web-based representation of 3d molecular structure as applied to proteopedia. Isr. J. Chem. *53*, 207–216.

98. Melanie, C. (2015). Burger. Chemdoodle web components: html5 toolkit for chemical graphics, interfaces, and informatics. J. Cheminf. *7*, 12.

99. Winther, K.T., Hoffmann, M.J., Boes, J.R., Mamun, O., Bajdich, M., and Bligaard, T. (2019). Osman Mamun, Michal Bajdich, and Thomas Bligaard. Catalysis-hub.org, an open electronic structure database for surface reactions. Sci. Data *6*, 75.

100. Github. https://github.com/.

101. C. Boettiger. An Introduction to Docker for Reproducible Research.

102. Pollice, R., Dos Passos Gomes, G., Aldeghi, M., Hickman, R.J., Krenn, M., Lavigne, C., Lindner-D'Addario, M., Nigam, A., Ser, C.T., Yao, Z., and Aspuru-Guzik, A. (2 2021). Data-driven strategies for accelerated materials design. Acc. Chem. Res. *54*, 849–860.

103. Alberi, K., Nardelli, M.B., Zakutayev, A., Mitas, L., Curtarolo, S., Jain, A., Fornari, M., Marzari, N., Takeuchi, I., Green, M.L., et al. (2019). The 2019 materials by design roadmap. J. Phys. D Appl. Phys. *52*, 013001. computational data is more homogeneous than experimental data as referenced in Himanen2019⟨br/⟩.

104. Ong, S.P., Richards, W.D., Jain, A., Hautier, G., Kocher, M., Cholia, S., Gunter, D., Chevrier, V.L., Persson, K.A., and Ceder, G. (2013). Python materials genomics (pymatgen): a robust, open-source python library for materials analysis. Comput. Mater. Sci. *68*, 314–319.

105. Openbabel (2016). The Open Source Chemistry Toolbox. http://openbabel.org/wiki/Main_Page.

106. Ghiringhelli, L.M., Carbogno, C., Levchenko, S., Mohamed, F., Huhs, G., Lüders, M., Oliveira, M., and Scheffler, M. (2017). Towards efficient data exchange and sharing for big-data driven materials science: metadata and data formats. npj Comput. Mater. *3*, 46.

107. Research Data Management Plan Guidelines.

108. Horton, M.K., and Woods-Robinson, R. (2021). Addressing the critical need for open experimental databases in materials science. Patterns *2*, 100411.

109. Coudert, F. (2019). Materials databases: the need for open, interoperable databases with standardized data and rich metadata. Adv. Theory Simul. *2*, 1900131.

110. Kearnes, S.M., Maser, M.R., Wleklinski, M., Kast, A., Doyle, A.G., Dreher, S.D., Hawkins, J.M., Jensen, K.F., and Coley, C.W. (2021). The open reaction database. J. Am. Chem. Soc. *143*, 18820–18826.

111. Jesper Jacobsson, T., Adam, H., and García-Fernández, A. (2021). An open-access database and analysis tool for perovskite solar cells based on the fair data principles. Nat. Energy *12*.

112. Xu, H., Thakur, K., Kamruzzaman, A.S., and Ali, M.L. (2021). Applications of cryptography in database: a review. In 2021 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS) (IEEE), pp. 1–6.

113. Microsoft Azure. https://azure.microsoft.com/en-us/.

114. Amazon. https://aws.amazon.com.

115. Del Valle, E. (2017). Sharing My Loss to Protect Your Data: A Story of Unexpected Data Loss and How to Do Real Preservation, *9*.

116. The Dataverse Project. https://dataverse.harvard.edu.

117. DataHub. https://datahub.csail.mit.edu.

118. Zenodo. https://zenodo.org.

119. NOMAD. https://nomad-coe.eu/.

120. Materials Project. https://materialsproject.org/.

121. AFLOW. https://aflowlib.org/.

122. Catalysis Hub. https://www.catalysis-hub.org.

123. JARVIS. https://jarvis.net.gov/.

124. Crystallography Open Database (COD). https://crystallography.net/.

125. Citrine Informatics. https://citrine.io/.

126. re3data. https://www.re3data.org.

127. Trusted Digital Repositories: Attributes and Responsibilities - An RLG-OCLC Report. Research Libraries Group; 2022. https://www.oclc.org/content/dam/research/activities/trustedrep/repositories.pdf;.

128. National Science Foundation. Where Discoveries Begin. http://www.nsf.gov/bfa/dias/policy/dmp.jsp.

129. EPSRC. https://www.ukri.org/councils/epsrc/.

130. Go FAIR. https://www.go-fair.org/.

131. Research Data Alliance. https://www.rd-alliance.org/.

132. CODATA. https://codata.org/.

133. FAIRsharing. https://fairsharing.org/.

134. Acceleration Consortium. Accelerating the Discovery of Materials and Molecules Needed for a Sustainable Future. https://acceleration.utoronto.ca/.

135. NOMAD. The FAIRmat Consortium. https://www.fairmat-nfdi.eu/fairmat/consortium.

136. University of Southern California (USC) Materials Consortium. https://matcon.usc.edu/.

137. Citrine Informatics. Unlocking the Power of Data in Materials and Chemical Development. Citrine Informatics

138. Musen, M.A. (2022). Without appropriate metadata, data-sharing mandates are pointless. Nature *609*, 222.