MaRDA FAIR materials microscopy and LIMS data working groups' community recommendations

Joshua A. Taillon,* Edward S. Barnard, Laura M. Bartolo, Maria K.Y. Chan, Eric A. Stach, Mitra L. Taheri, L. Catherine Brinson, and Peter W. Voorhees

Received: 23 November 2024 / Revised: 27 January 2025 / Accepted: 5 February 2025

Managing, processing, and sharing research data and experimental context produced on modern scientific instrumentation all present challenges to the materials research community. To address these issues, two MaRDA Working Groups on FAIR Data in Materials Microscopy Metadata and Materials Laboratory Information Management Systems (LIMS) convened and generated recommended best practices regarding data handling in the materials research community. Overall, the Microscopy Metadata Group recommends (1) instruments should capture comprehensive metadata about operators, specimens/samples, instrument conditions, and data formation; and (2) microscopy data and metadata should use standardized vocabularies and community standard identifiers. The LIMS Group produced the following guides and recommendations: (1) a cost and benefit comparison when implementing LIMS; (2) summaries of prerequisite requirements, capabilities, and roles of LIMS stakeholders; and (3) a review of metadata schemas and information-storage best practices in LIMS. Together, the groups hope these recommendations will accelerate breakthrough scientific discoveries via FAIR data.

Impact statement

With the deluge of data produced in today's materials research laboratories, it is critical that researchers stay abreast of developments in modern research data management, particularly as it relates to the international effort to make data more FAIR – findable, accessible, interoperable, and reusable. Most crucially, being able to responsibly share research data is a foundational means to increase progress on the materials research problems of high importance to science and society. Operational data management and accessibility are pivotal in accelerating innovation in materials science and engineering and to address mounting challenges facing our world, but the materials research community generally lags behind its cognate disciplines in these areas. To address this issue, the Materials Research Coordination Network (MaRCN) convened two working groups comprised of experts from across the materials data landscape in order to make recommendations to the community related to improvements in materials microscopy metadata standards and the use of Laboratory Information Management Systems (LIMS) in materials research. This manuscript contains a set of recommendations from the working groups and reflects the culmination of their 18-month efforts, with the hope of promoting discussion and reflection within the broader materials research community in these areas.

Introduction

Until recently, debate on the findable, accessible, interoperable, and reusable (FAIR) data principles in materials research focused largely on whether to support and promote its adoption.[1] Efforts related to the adoption of the FAIR principles in materials science have been increasing in recent years and are international in scope. For example, in 2023, a one-day workshop in Berlin emphasized the need and proposed shared metadata measures in the materials sciences.[2] Separately, a body of recognized materials experts in the United States came together to advocate for specific actions that needed to be undertaken by the materials community at large and by individual researchers within the community.[3] The collective preliminary work endorsed both the materials community's involvement in defining subfields of materials research, such as materials microscopy, as well as individuals' roles to plan, prepare, and submit their research data in order to assemble

Joshua A. Taillon, Material Measurement Laboratory, National Institute of Standards and Technology, Boulder, USA; joshua.taillon@nist.gov
Edward S. Barnard, Lawrence Berkeley National Laboratory, Berkeley, USA; esbarnard@lbl.gov
Laura M. Bartolo, Center for Hierarchical Materials Design, Northwestern University, Evanston, USA; laura.bartolo@northwestern.edu
Maria K.Y. Chan, Center for Nanoscale Materials, Argonne National Laboratory, Lemont, USA; mchan@anl.gov
Eric A. Stach, Department of Materials Science and Engineering, University of Pennsylvania, Philadelphia, USA; stach@seas.upenn.edu
Mitra L. Taheri, Department of Materials Science and Engineering, Johns Hopkins University, Baltimore, USA; mtaheri4@jhu.edu
L. Catherine Brinson, Department of Mechanical Engineering and Materials Science, Duke University, Durham, USA; cate.brinson@duke.edu
Peter W. Voorhees, Department of Materials Science and Engineering, Northwestern University, Evanston, USA; p-voorhees@northwestern.edu
*Corresponding author
doi:10.1557/s43577-025-00882-2

significant amounts of FAIR materials data and enable breakthrough materials research.

With the growing prevalence of artificial intelligence (AI), deliberations have taken a considerable shift to concentrate more directly on how best to implement FAIR data into materials research practices quickly and efficiently to turn AI's benefits into high-impact discoveries in materials research.[4] In 2022, the National Science Foundation launched a particularly effective collaborative effort through its Findable, Accessible, Interoperable, Reusable, Open Science Research Coordination Networks (FAIROS RCNs) programs to establish Research Coordination Networks in critical fields and geographic regions.[*] Through FAIROS, the Materials Research Coordination Network (MaRCN) was established to enable the Materials Research Data Alliance (MaRDA)[5] to accelerate connection across the materials research community through activities needed to create and utilize FAIR data. To support open-science materials research nationally and internationally, MaRCN aims to bridge the fundamental gap between materials data and data-intensive methods, including AI and machine learning (ML). The MaRCN project involves six institutions: Johns Hopkins University (the lead institution); Duke University; Northwestern University; Purdue University; University at Buffalo, The State University of New York; and The University of Chicago. One focus of MaRCN – FAIR DATA – was led by Northwestern University and Duke University to host activities for academic and industry researchers aimed at fostering concurrent development of recommended best practices to describe and manage materials data.

Since their publication in 2016,[6] the FAIR data principles have been adopted, implemented, and adapted into scientific practices across the science domains with varying yet increasing degrees of success and endorsement. In health research, an open architecture workflow process transformed raw, unorganized health data by following the "GO FAIR" "FAIRification" process resulting in identified gaps of the process for reusable health data sets.[7] An open-access database and analysis tool for perovskite solar cells based on published research and following the FAIR data principles has been developed and made publicly accessible with applicability to materials science, engineering, and biosciences.[8] For the field of tribology, work recognizing the value of FAIR data and the lack of community-developed and accepted methods to describe tribological experiments, has laid out needed documentation to incorporate the principles into practices.[9] Drug R&D in a biopharmaceutical private/public enterprise were implemented incorporating FAIR data principles in 2019[10] and with more recent combination of AI and FAIR data to advance drug discovery.[11]

To support the MaRCN goals, Northwestern University and Duke University, as members of the MaRDA Advisory Council, jointly held a Virtual Materials Community Meeting on December 8, 2022 with more than 100 attendees (primarily drawn from the United States and MaRDA membership) and invited presentations in two high-priority areas given by leading experts in the fields:

- Mitra Taheri[†]
- June Lau[‡]

Considerable data challenges in materials science result from data generated by electron microscopes, given their near ubiquitous presence in every materials department, national laboratory, and industry as well as the need for data sharing across multiple organizations with varying capabilities. One solution to address these challenges is the adoption of Laboratory Information Management Systems (LIMS) to support the production, capture, and management of highly heterogeneous, large-scale data sets. Integrating LIMS strategies into materials data workflows has been limited, however, by lack of awareness and expertise within the materials research community. Bringing together the dual foci of materials microscopy data and LIMS melds the individual impacts of a widely used experiment tool with a data-life-cycle framework applicable across materials research with the potential to deliver great community benefits.

At the December 2022 Virtual Materials Community Meeting, MaRCN extended invitations to all meeting participants to contribute to the establishment of two MaRDA working groups (WGs) in these key, complementary areas important to the materials research community: (1) materials microscopy metadata and (2) LIMS. MaRDA WGs are 18-month long community-led efforts to establish community best practices, advance data sharing, and spur innovation.

In January 2023, Northwestern University and Duke University established these two MaRDA WGs with co-chairs and members comprised of recognized materials leaders and experts in the areas of materials microscopy and LIMS. The materials microscopy data WG focused on defining high-impact community data generation best practices for materials microscopy metadata while the LIMS WG addressed best practices for individuals to plan, prepare, and complete the integration of LIMS into materials research data

[*]US National Science Foundation Findable Accessible Interoperable Reusable Open Science Research Coordination Networks (FAIROS RCN) NSF 22-553 https://www.nsf.gov/pubs/2022/nsf22553/nsf22553.htm supports this portion (NSF FAIROS RCN: 2226417) of the Materials Research Coordination Network as part of NSF's RCN program to advance and coordinate findable, accessible, interoperable, reusable (FAIR) data.

[†]https://orcid.org/0000-0001-5349-1411. (Johns Hopkins University) presented *Microscopy and FAIR Data*

[‡]https://orcid.org/0000-0002-5233-4956. (National Institute of Standards and Technology–NIST) presented *Electron Microscopy Facility Data Management: NexusLIMS (Laboratory Information Management Systems)*

management workflows. Both groups concentrated their efforts on the types of "nonvalidated" environments typical in the experience of WG members (i.e., academic, government, and noncertified industrial research environments). In certified research or testing environments, stricter requirements are explicitly defined by a number of standards such as ISO 17025[12] and ISO 9001[13] that provide specific guidance to meet the needs of these laboratories. The materials microscopy WG was led by co-chairs Edward Barnard (Lawrence Berkeley National Laboratory), Maria Chan (Argonne National Laboratory), and Mitra Taheri (Johns Hopkins University) with 10 members.[§] The LIMS working group was led by co-chairs Eric Stach (University of Pennsylvania) and Joshua Taillon (NIST) and 10 members[**] with MaRDA Advisory Council members Laura Bartolo (NU), Cate Brinson (Duke University), Peter Voorhees (NU), and June Lau (NIST) as ex officio members of both WGs.[††]

While the MaRDA WGs on materials microscopy metadata and LIMS were separate entities and followed independent processes, they were closely related and both supported by the MaRCN staff. Each group's stated goal was not to develop novel approaches or techniques, but rather to review current approaches within each groups' remits and present a set of approachable recommendations to those in the community that are not experts in data science or data management. To bring their synergistic efforts together for increased opportunities of exchange, adoption, and broad community impact, Northwestern University hosted two joint, in-person, 1.5-day meetings for both WGs in May and October 2023. Each WG independently held multiple additional virtual meetings to conduct and build upon their efforts during the intervening 18-month period. Preliminary draft reports and requests for feedback were presented at the 2024 MaRDA Annual Meeting (Virtual, February 22, 2024), the Midwest Microscopy and Microanalysis Meeting (Northwestern University, March 15, 2024), and the 2024 Spring MRS Meeting (Seattle, April 22, 2024)[14] and posted online.[5] These two MaRDA WGs formally concluded their efforts in October 2024 and now present their respective recommended best practices in this article.

Recommendations from the materials microscopy working group

One of the largest and fastest growing data challenges in materials science is data generated by microscopes. These instruments are present in nearly every materials science and engineering department, national laboratory, and many industries, making it challenging to reach consensus on critical metadata and ontologies as well as to facilitate data sharing both intramurally, as well as across multiple organizations with varying capabilities. Additionally, with the growing prevalence of AI and machine learning techniques there is a need to aggregate microscopy data and metadata in a consistent manner to aid in the training of such ML models.

As a foundational step toward recommended minimal, common, lightweight metadata for materials electron microscopy, the MaRDA materials microscopy WG surveyed the landscape of electron microscopy metadata standards and metadata practices in cognate disciplines (e.g., life sciences, materials science, and chemistry). Many scientific communities have attempted to tackle the problem of data standardization in the hopes of enabling FAIR data sharing. This includes development of common data formats as well as shared naming schemes or ontologies to ensure that there is consistent meaning to a quantity across scientists, instrument vendors, and subcommunities. Here, we highlight some examples of such efforts and the lessons we can learn from them. It should also be noted that in addition to the formal approaches outlined next, ML techniques are starting to assist in the generation of metadata standards themselves through natural language processing of the corpus of materials research literature.[15]

OME-XML: The Open Microscopy Environment (OME) is an open-source software framework and community-driven initiative that aims to support the exchange and analysis of biologic microscopy data.[16] It provides tools and resources to enable researchers to manage, share, and analyze large sets of (primarily biologic) microscopy images efficiently. OME emphasizes open standards, creating a flexible infrastructure that can accommodate various imaging modalities, file formats, and metadata standards. It is targeted predominantly at optical microscopy for biology applications. OME includes standards for metadata representation, such as OME-XML, which provides a structured way to describe the acquisition parameters, instrument settings, and sample details associated with microscopy images. It enables the description of various aspects, such as acquisition parameters, instrument settings, and sample details, ensuring comprehensive documentation of experimental conditions. Because of this focus, it includes standard naming conventions for optical components, such as "Filter," "Objective," and "Laser."

[§]MaRDA Materials Microscopy Metadata WG Members: Eva Campo (Campostella Research), Fernando Castro (Gatan Inc.), Miaofang Chi (Oak Ridge National Laboratory), John Damiano (Protochips Inc), Anthony DiGiovanni (Army Research Laboratory), Tom Isabell (JEOL), Robert Klie (University of Illinois at Chicago), Jia Ying (Northwestern University–NU), Prashant Singh (Ames National Laboratory), Maureen Williams (NIST).

[**]MaRDA LIMS WG Members: John Allison (University of Michigan), Carelyn Campbell (NIST), Jennifer Carter (Case Western Reserve University), Kamal Choudhary (NIST), Cory Czarnik (Gatan Inc), Dieter Isheim (NU), Derk Joester (NU), Roberto dos Reis (NU), Richard Sheridan (Duke University), Douglas Stauffer (Bruker Corporation).

[††]Although international perspectives are of critical importance in forming broad consensus within the community, funding requirements of the NSF FAIROS program limited participation in the working groups to the materials research community within the United States.

NeXus and NXem: NeXus is an open format for the storage and exchange of scientific data, commonly used in neutron, x-ray, and muon experiments.[17] The NXem draft extension to the NeXus file format is specifically designed to capture the data and metadata from electron microscopy imaging and spectroscopy.[18] Due to the accelerator focus of the NeXus standard, the NXem extension describes EM as an "electron accelerator" and its naming conventions follow this logic. In an effort related to NXem, the Helmholtz Metadata Collaboration has published an *Electron Microscopy Glossary*,[19] which provides a community-curated formal vocabulary for terms commonly used in EM (and provides definitions of the terms used in the NXem NeXus extension). A formal vocabulary such as this can serve as a "semantic clearing house" and be used to unequivocally indicate (in a machine- and human-readable way) the meaning of metadata terms, regardless of the specific metadata format used.

HMSA – Hyper-Dimensional Spectral Data File Format: The Hyper-Dimensional Data File Specification (HMSA) is a standard developed in collaboration with the Microscopy Society of America (MSA), the Microanalysis Society (MAS), and the Australian Microbeam Analysis Society (AMAS) for the exchange of hyper-dimensional microscopy and microanalytical data between different software applications.[20] There is a clear focus on electron microscopy techniques that include traditional imaging modalities along with spectroscopic and diffraction techniques. The format has been standardized via the ISO standardization process as ISO 5820. HMSA data sets consist of a pair of files: An XML text document for metadata and an uncompressed binary file to store raw data. The metadata file contains "conditions" of the instrument at the time of data acquisition. These categories include "Instrument," "Probe," "Specimen," "Specimenenvironment," "Measurement mode," "Detector," "Acquisition," and "Calibration." Additional details in each category are well defined in the specification for different instrument types (i.e., SEM, TEM) and measurement modalities.

Materials microscopy WG: Recommended best practices

The Materials Microscopy WG has developed a set of recommended best practices for managing and utilizing metadata in materials microscopy. These guidelines are designed to enhance data quality, interoperability, and reproducibility in materials research, particularly in the realm of electron microscopy. By following these best practices, researchers can ensure that their microscopy data are well-documented, easily accessible, and valuable for future studies, including the future of AI-driven data analysis.

Comprehensive metadata capture

In general, more metadata are better. The complete capture of the context of a data set should be represented in its metadata with an effort made to standardize naming and organization of this information (see **Figure** 1). However, these standards should not constrain what metadata is included. Additional labeled metadata fields should be included, and the data format used should be extensible enough to allow for unlimited extra fields.

We categorize desired metadata into four categories: (1) core bibliography information, (2) specimen/sample information, (3) instrument conditions, and (4) image data information:

1. The core bibliographic information includes answers to the questions: Who? What? Where? When? Basic bibliographic data such as these can be encoded using Dublin Core standards.[21]
2. Sample information should be complete enough to uniquely identify the sample and its process through the use of persistent identifier,[22] (i.e., not merely "Sample A," but rather a full description [such as an IGSN, for example, 10.58151/NHB00377H]).[23]
3. Microscope conditions should include the full information needed to replicate the measurement. For example, in a TEM this should include information, including accelerating voltage, magnification, camera length, and defocus. See **Table** I next.
4. Image data metadata should include technical information that describes the data and their formatting. This includes file type, imager information, gain settings, and pixel sizes. Additionally, the format of the data file should be fully defined and preferably in an open format such as TIFF[24] or HDF5.[25]

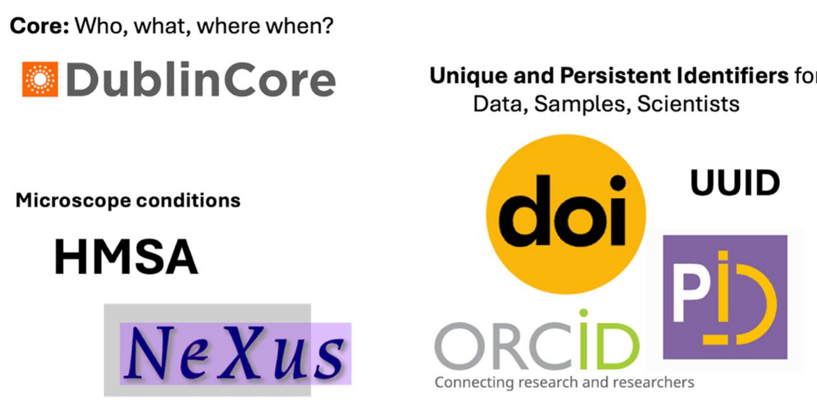

Figure 1. Examples of existing metadata and identifier standards that can be used by materials EM community. For more information on each, please refer to the discussion in the text and consult the following references: Dublin Core,[21] HMSA,[20] NeXus and NXem,[17,18] DOI,[22] ORCiD,[29] UUID,[30] and PIDInst.[31]

Table I. Examples of metadata (nonexhaustive) that should be included and what they should be called. Where possible, a formal vocabulary for terms (such as Reference 19) should be used. For more examples, please consult References 18–20.

Metadata Term	Recommendation	
Data Set Identifier	UUID for all data, DOI for published/curated data	
Microscope Name and Model	Follow the PIDInst standard[31]	
Instrument Unique Identifier	PIDInst (e.g., 21.T11998/0000-001A-3905-F)	
User Unique Identifier	Name, Email, and ORCiD (e.g., 0000-0003-4736-0743)	
Sample/Specimen	Full description, such as an IGSN (e.g., 10.58151/NHB00377H)	
Microscope Conditions (for a TEM example)	**HMSA Standard[20]**	**NeXus NXem[18]**
Beam Current	⟨BeamCurrent Unit = "nA"⟩	NXoptical_system_em/beam_current
Accelerating Voltage	⟨BeamVoltage Unit = "kV"⟩	NXebeam_column/electron_gun/voltage
Magnification	⟨NominalMagnification⟩	NXoptical_system_em/magnification
Camera Length	⟨NominalCameraLength Unit = "cm"⟩	NXoptical_system_em/camera_length
Defocus	⟨Defocus Unit = "nm"⟩	NXoptical_system_em/defocus
..	..	..

Dublin Core metadata is designed to provide a simple and standardized way to describe digital resources such as documents, images, web pages, and other types of content. The Dublin Core Metadata Initiative (DCMI) developed and maintains this set of metadata terms, aiming to improve the discoverability, accessibility, and management of digital resources.[21] The Dublin Core Metadata Element Set includes 15 core elements, each represented by a term and accompanied by a definition. These elements cover basic descriptive information about a resource, such as title, creator, description, contributor, and date that are relevant to all documents, including microscopy data.

Metadata should aim to document all relevant experimental conditions, including sample preparation methods, microscope settings (e.g., accelerating voltage, magnification, and detector type), and environmental conditions (e.g., temperature and vacuum levels). It should record any deviations from standard protocols to provide context for the resulting data, as well as maintain detailed records of instrument calibration procedures and results. This includes calibration of the electron source, lenses, and detectors. Additionally, it is critical to automate collection of metadata as much as possible, as human error or inaction will lead to uncertain or missing information that cannot be recovered later. Data acquisition software should provide easy and obvious ways to record such metadata, and should record metadata into produced data files in a consistent and open manner.

Unique persistent identifiers (PIDs) are crucial for scientific data as they ensure long-term accessibility and traceability of data sets, facilitating reproducibility and verification of research findings. Microscopy metadata should include links to other relevant data sources (such as a lab notebook entry, sample tracking database, and instrument database) and unique identifiers and canonical persistent links are important in maintaining reliable connections between data. For an accessible introduction to PIDs, see Reference 22.

Recording and standardizing data units in metadata are critical for accurate analysis. Without a proper understanding of units and normalization, computers cannot accurately process data.[26] The use of a consistent unit of measurement, such as electron volts or kilo electron volts, is suggested to facilitate conversion and analysis, and these units should be represented in a standardized way. Furthermore, the normalization of data is underscored as crucial to ensure consistency and accuracy in analysis. In general, the materials microscopy community should follow the recommendations of the Digital Representation of Units of Measurement (DRUM) Task Group.[27]

With the current push to develop "digital twins" of instruments – models that can simulate the entire instrument that are verified and updated by experiments – metadata has an important role in providing enough information that the data can be reproduced in a model. Thus, a laudable stretch goal of metadata is to provide this full context.[28] For this to be possible, close collaboration with microscope vendors is key, as they are best positioned to develop such digital twins and provide insight into the needed metadata for reconstruction of data.

Standardized metadata schema

We recommend employing controlled vocabularies and standardized terminologies to describe microscopy data and metadata. This promotes consistency and facilitates data sharing and comparison across different studies and laboratories. However, in our survey of existing standards, there are many examples of terminology and schema that meet the needs of specific scientific domains or use cases. It is unlikely that there is truly a single schema that can efficiently capture the needs of all scientists, but minimizing the number of standards is reasonable (see **Figure** 2). What is then also needed is efficient transformation of metadata standards from one to another. As previously mentioned, ML techniques are beginning to make

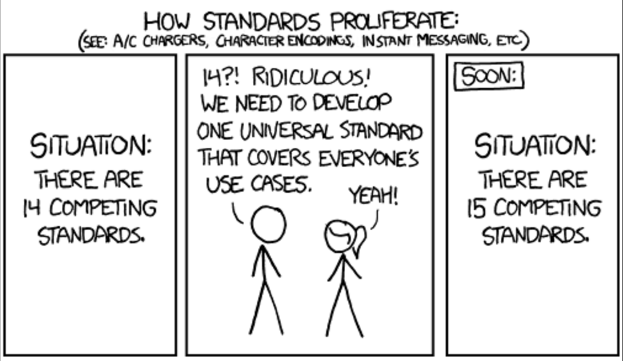

Figure 2. A (humorous) perspective on how the desire to improve existing standards can lead to an endless proliferation of standards, making them all less effective in the process.[32] We appreciate that materials microscopy metadata is susceptible to this same effect, and as such advocate for community adoption of consensus standards as they are developed (such as those identified in this work) rather than the wholesale development of new standards.

strides in defining ontologies and could soon provide automated translation between standards.[15]

With that said, we have identified standards that capture the needs of the materials electron microscopy community. For bibliographic information, the Dublin Core can provide answers to who, what, where, and when. For microscopy conditions we have found HMSA, now an ISO standard (ISO/DIS 5820 under development),[20] and the NXem extension to the NeXus file format provide an effective ontology for describing materials EM instrument conditions.[18] Thankfully, community-developed tools currently exist (such as *RosettaSciIO*[33]) to translate the vast array of proprietary, and often closed, file formats into a common open format, although work is needed on consistent metadata definitions. Finally, unique identifiers and persistent links can be provided through standards and organizations such as ORCiD for uniquely identifying people,[29] PIDInst (Research Data Alliance PID for instruments),[31] and DOIs or other PIDs for data sets.[22] Other globally unique identifiers can also be generated using the Internet Engineering Task Force (IETF) proposed UUID standard for other elements not handled by the previously identified standards.[30] As an example of using these standards, see Table I.

Recommendations from the LIMS WG

The MaRDA LIMS WG was convened to bring together experts in materials science from a range of backgrounds. Its primary aims were to evaluate the current state of the art and generate a set of actionable recommendations for the community to facilitate the adoption of LIMS throughout materials research. The WG included representatives from academia, government laboratories, and industry partners, illustrating the wide-ranging interest in and recognition of the importance of modernizing laboratory data handling in the materials community. Although simple data curation strategies (such as organizing data into folder hierarchies and embedding metadata in filenames) could work for individual researchers or small research teams, such bespoke approaches quickly limit interoperability in an ever more interconnected modern research environment. Thus, coordination at the community level (such as through MaRDA WGs) is necessary to better promote the generation of materials data conforming to the FAIR principles.[6,34]

At the outset of the WG's efforts, the members all agreed that laboratory information management is an essential component of modern materials research laboratory operations and can provide the digital infrastructure necessary to support a range of essential services, including data management, sample tracking, and reporting of results. It quickly became evident, however, that interpretations of the term LIMS can (and do) vary greatly within our community. Thus, throughout the WG's efforts, we adopted the definition used in *NIST Technical Note 2216* of LIMS as "a system of components, which delivers the capabilities for the early stages of a research life cycle" (Sec. 4 of Reference 35). This definition acknowledges there is no "singular LIMS solution" ideal for all use cases but envisions a LIMS as an interconnected network of composable components using standardized practices, allowing for a lower barrier for entry and ensuring scalability.

In discussing LIMS, this document focuses on those directly involved with implementing and managing the LIMS within a laboratory or group of integrated laboratories. Although additional relevant participants and stakeholders from within an institution (such as research administrators, librarians, professional organizations, funding agencies, grant and program officers, and the public) are not discussed in detail here, this document recognizes their expertise and that their interests are important. Together with those directly involved with laboratories, they represent essential stakeholders in a fully developed and accountable data curation, management, and publication system in order to implement established FAIR standards and best practices.

To best promote the adoption of LIMS tools within the materials community, the group decided to focus on three key areas: (1) an analysis of the tradeoff of costs and benefits involved in implementing a LIMS; (2) an investigation of what prerequisites are required to implement a LIMS and what capabilities it enables for various roles in the system; and (3) an introduction to and recommendations about metadata schemas and best practices to be used to catalog information within a LIMS for materials research. These topics were identified during initial WG meetings as areas where all members agreed there was a current lack of clarity, as informed by discussions with colleagues from the materials community. Thus, a primary goal of the WG was to try to provide a materials researcher (who is likely not a data management expert) with the tools necessary to evaluate their current research data management environment, identify areas for improvement, and devise an actionable plan to implement the portions of a LIMS that would enhance their research data workflows. It is the intent of the WG that these recommendations stand in addition to (and not in place of) the discussion of Reference 35.

Costs and benefits of LIMS

Although a researcher may already understand the advantages of integrating a LIMS into their workflow, individuals are likely limited in their ability to implement substantial changes or make additions to the digital infrastructure of their organization. This could include, for example, the use of digital scheduling systems, centralized data storage, automated data transfer solutions, electronic laboratory notebooks (ELN), etc., which cannot be unilaterally adopted in the sort of shared infrastructure common in today's shared research environment. Because of this, it is crucial to obtain buy-in from higher levels of the organization, which typically comes down to a cost/benefit analysis that is, "What do we have to spend, and what will we get for it?"

It is important to acknowledge that the costs borne and benefits realized from a LIMS will differ depending on a person's role in an organization and that costs are not solely financial in nature. For example, a research group leader or department chair will be likely interested in the financial costs related to acquiring software or storage hardware, paying salaries for system maintenance, etc., whereas an individual researcher will be concerned with the related (short term) cost of reduced productivity while learning new systems and adjusting their workflow to new approaches. Both types of costs can contribute to hesitancy from across an organization and it is critical to be sensitive to the needs of all stakeholders when proposing changes. Where at all possible, the "intangible" costs to individual researchers should be minimized by adapting to existing procedures in order to promote positive engagement with a new LIMS system.

Through a review of relevant literature, internal discussions, and interviews with various materials research facility managers, the WG identified several benefits to be realized from the adoption of a LIMS in a materials research environment. At a group leader or organizational level, a primary benefit is to enhance data discovery and promote collaboration using standardized, searchable, and machine-readable data and metadata. This in turn improves the reliability of research results and experimental reproducibility, easing compliance with the FAIR data principles, which are increasingly seen in funding agency requirements. Furthermore, making data easily machine readable will allow it to be better utilized in automated data analysis routines and as a data source for various ML approaches, including large language models (LLMs) combined with retrieval-augmented generation (RAG).[36] At the individual researcher level, such systems can remove frequent burdens, such as organizing data, maintaining backups, and correlating files with their metadata (often in a laboratory notebook). A LIMS can additionally provide rapid access to historic notes and data and simplify the sharing of such data with collaborators. These benefits collectively allow the researcher to dedicate more time to the research process itself, rather than the "overhead" of individualized data management practices. For further discussion of the benefits afforded by LIMS and ELN platforms, see References 37 and 38.

At a financial level, interviews with various facility managers revealed a range of monetary costs associated with LIMS deployment, depending on the complexity and scale of the solution chosen. For an individual research group, the cost could feasibly be as low as USD$10,000 when accounting for a solution powered by consumer-grade network attached storage, open-source software, and the part-time labor of a graduate student. At an institutional scale, the most cited figures indicated a one-time cost in the range of approximately USD$30,000 to $150,000, depending on the need to procure storage hardware, decisions of on-site versus cloud storage, etc. There is no strict upper limit to these costs, as they will scale with the amount of storage (or redundancy) needed, but these figures were the typically cited range in our interviews. It is important to acknowledge that there will be ongoing costs as well for the maintenance of a system, requiring an institutional commitment. These ongoing requirements could include (1) refreshing or expanding data storage as needed, (2) staffing to maintain the system, (3) staff costs to train and support new and existing users, and (4) any potential license/subscription fees for software – if using a commercial solution. Hardware upgrade costs will depend on the amount of storage, but will be intermittent (perhaps every few years). Once a system is in place, ongoing staffing costs likely range from a fraction to half of a full-time employee (FTE), but costs related to support and training should decrease over time as LIMS becomes a known and integrated part of facility operations. It should be noted that in the WG's experience, low-end initial LIMS investments can often lead to increased long-term costs (when compared to a more extensive initial solution) due to a lack of resiliency, documentation, testing, etc., and this balance should be weighed carefully during the planning process. The WG recommends that those in the decision-making process consider expenses related to data management as important as those that are readily incurred for physical equipment.

Ultimately, it cannot be overstated that the WG's research revealed that financial considerations/costs are usually not a primary barrier (e.g., at a major facility, total LIMS costs would likely be under 1% of total expenditures). Rather, it is the challenge associated with being able to identify and hire personnel with the correct skill sets to implement and maintain a LIMS, as well as receiving buy-in from administration to prioritize the associated expenses as is readily done for research instrumentation.

LIMS roles, prerequisites, and capabilities

A LIMS can provide a range of capabilities that are essential components of modern laboratory operations. Their implementation, however, is not a simple task. It requires considerable planning and preparation, together with an understanding of the available technological solutions, the laboratory's needs, and the organization's overarching goals. As mentioned previously, a LIMS can be implemented at many different levels within an organization, for example, an individual laboratory, a group of research facilities, or coordinated across an entire organization in academia, government, or industry. Throughout this work, the general term "organization" is used to refer to a group implementing LIMS at any such level. While the scale of need will change with the scale of deployment, the fundamental requirements remain quite similar. As such, the WG has developed a recommended set of roles, prerequisites, and capabilities to serve as a concise and actionable reference for those seeking to implement LIMS solutions in their organization. The recommendations described next are also supplied in a convenient "checklist" style format in the supplementary materials.

Roles

Using the NIST Research Data Framework[39] as a guide, the WG has identified a matrix of the most important stakeholder roles for LIMS within an organization, and the activity topics for which each role should have primary or secondary responsibilities. Identifying individuals to serve in each of these roles can help to bring together a project team that is most responsive to the needs of the organization and make a LIMS deployment as impactful and beneficial as possible. At a high level, five primary stakeholder roles were identified: Researcher, Facility Manager, Data Manager, IT Manager, and Instrument Vendor/Product Manager. Each of these roles should have input related to LIMS planning, data/metadata generation, and data processing/analysis. For further detail on the specific recommended responsibilities of each role, please refer to the Supplementary materials.

Prerequisites

During the initial planning stages, prior to a LIMS implementation, it is critical for an organization to bring together a project team with representatives from all relevant departments and stakeholders to address the roles previously recommended, such that this group can clearly define the goals and objectives of the LIMS deployment (e.g., "What new functionality does this team need from the new system?"). The project team must have a thorough understanding of the laboratory's existing research workflows and processes, together with a comprehensive inventory of all lab equipment, instruments, and software (including knowledge about what file formats are produced). As the group progresses closer toward implementation, a plan for data migration from existing systems (if any) and disaster backup/recovery plans and policies should be specified. To ease the burden felt by users of the new system, a plan for user onboarding, ongoing training, and support needs to be developed, together with a plan for ongoing maintenance, testing, and system updates.

Capabilities

As a LIMS is typically a system of interconnected components,[35] the WG suggests a LIMS implementation should be modular and provide as many of the following technical capabilities as possible. With a modular approach, a LIMS deployment can progress in a piecemeal fashion, enabling new features as resources allow. The following capabilities are considered essential "core" features for a functional LIMS:

- Centralized automated collection and storage of research data and metadata
- Access rights control to limit access to centralized research data as necessary
- Use of persistent identifiers (PIDs) wherever possible within the LIMS (e.g., Handles, ARK IDs, DOIs, IGSNs, etc.)[22]
- Data and metadata within the LIMS are searchable for later retrieval and analysis
- Interfaces with instrument scheduling and laboratory facility management software (if present)
- All components of the LIMS provide and can consume data via well-documented application programming interfaces (APIs)

A complimentary set of capabilities were identified as recommended (but perhaps not essential):

- Data and metadata collection integrates with existing and novel research workflows (such as the use of an ELN)
- Data and metadata follow recommended standardized schemas (see the section "Metadata schemas for LIMS")
- Experimental data are stored in, or system provides automatic conversion to, open data formats (making use of tools, such as Reference 33)
- Supports the creation of derivative data and data metrics, within the LIMS or by integration with external tools
- Integrates or supports external data publication repositories
- Interoperable with other laboratories and LIMS systems to support data exchange (open and well-documented API layers)

Metadata schemas for LIMS

In addition to the previously discussed recommendations about the planning and design of a LIMS implementation, the WG also aimed to provide recommendations to the materials research community related to the types and structure of information that should be captured by a LIMS to maximize its utility. To this end, the WG reviewed and analyzed a number of data models used by related projects, from both within and outside of the materials community. These included general-purpose data schemas, including the Dublin Core Metadata Initiative,[21] Schema.org,[40] and the Data Catalog Vocabulary (DCAT),[41] as well as the metadata models behind general data repository tools, such as Figshare[‡‡,42] and the Open Science Framework.[43] Additionally, the WG evaluated materials and other scientific data-specific schemas such as those published by the Materials Data Facility,[44] Foundry-ML,[45] the NexusLIMS project,[46,47] and the Sandia National Laboratories' Ecosystem for Open Science,[48] itself an extension of DCAT, named DCAT-eOS-AP. Although we acknowledge this list of data models is nonexhaustive, we limited our analysis to the above list during the efforts of the WG.

After a thorough review of the aforementioned schemas, the WG recommends a LIMS metadata structure that has the following characteristics:

- **Basic information:** The schema must allow for easy storage and recall of basic information about data, such as:
  - Who collected it (e.g., a researcher, technician, and assistants)
  - What are the data (type, sample relationship, etc.)
  - When/where was it collected (physical location and originating instrument)
  - Ideally, the system will allow for contextual information as well about why the data were collected
- **Data organization:** The "core organizing unit" of the schema should be a `Dataset`:
  - `Datasets` can consist of one or more individual `Files`
  - Metadata related to an experiment can be defined at both the `Dataset` and `File` levels; allowable metadata can change depending on the type of `File`
  - The schema should allow for explicit definitions of `Projects` as a way to indicate relationships between various components
  - `Datasets` can be composed into higher-order conceptual groupings (e.g., an `Experiment`, a `Run`, a `Collection`, or any other grouping as applicable to a domain)
- **Extensibility:** The number of required fields should be kept to a minimum:
  - Enforcing a minimal core metadata model provides the most utility to the widest group
  - Allowing optional granular metadata parameters enables a rich expression of experimental context unique to individual subdomains (for example, a microscopy metadata standard – see the section "Standardized metadata schema")
- **Linking and interoperability:** Existing community standards should be used:
  - Where feasible, standard (ideally persistent) identifiers should be used throughout the system (e.g., samples referenced by IGSN,[23] instruments by PIDInst identifiers,[31] people by ORCiDs,[29] and organizations by RORs).[49]
  - Items created within the system (`Datasets`, `Files`, etc.) should have persistent identifiers created along with them (such as a Handle and ARK ID) – this could necessitate deployment of or subscription to a service to create PIDs [50,51]
  - The LIMS metadata model does not need to be monolithic; rather, it should be interoperable and allow for linkages with other more domain-specific schemas, such as for samples, materials, and processes.

Of the schemas evaluated by the WG, the DCAT-eOS-AP data model[48] most closely adheres to the previously mentioned recommendations. As an example, it splits the data storage model into two primary groups (illustrated in **Figure** 3): "core" elements that pertain to any type of data and "granular" elements that contain domain-specific metadata. The DCAT-eOS-AP model is extensible, as other domains can plug into the model with customized schemas at the granular level. Although the

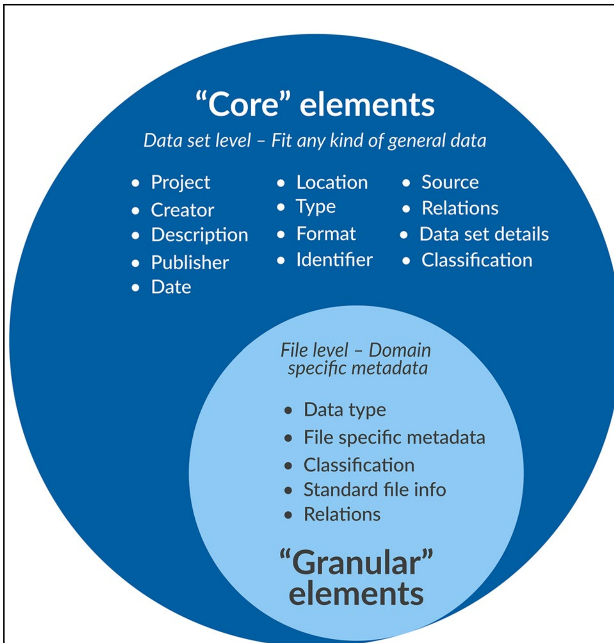

Figure 3. An example of an extensible metadata structure split into "core" and "granular" elements from the Data Catalog Vocabulary-Ecosystem for Open Science-Application Profile metadata model, adapted from Reference 48.

‡‡Certain commercial products and vendors are identified in this work for context and informational purposes. Such identification is not intended to imply recommendation or endorsement by NIST, nor is it intended to imply that the products identified are necessarily the best available for the purpose.

WG acknowledges it is likely not perfectly suitable for every use-case "out of the box," the WG endorses its design and suggests it as a strong starting point for a LIMS schema for the materials community.

Conclusion

MaRDA WGs are 18-month community-driven efforts aimed at accelerating progress in data-driven innovation through data sharing, exchange, and interoperability. The separate yet complementary endeavors of the WGs on materials microscopy and LIMS brought together stakeholders from universities, government laboratories, and industry partners for substantive in-person and virtual discussions focused on achievable, outcome-oriented best practices in two central data challenges facing the materials research community. The near-term impact of the WGs' accomplishments in materials data use and reuse lays a solid foundation to continue progress on building consensus for future FAIR data enhancements.

In providing these recommendations to the materials research community, the WGs hope to spur continued discussions within the community, and raise awareness of how both efforts can benefit individuals, institutions, and the community as a whole through better practices in the recording and sharing of materials data. While truly there is no "one-size-fits-all" solution, the WGs believe that systems composed of reusable modular pieces stand the best chance of bringing modern data management practices to the materials research community. We welcome feedback and further discussion of the recommendations presented in this work and hope they could inspire individuals and organizations throughout our community to seriously consider incorporating formal microscopy metadata and LIMS approaches into their research data workflows. Furthermore, although the scope of the WGs' efforts did not allow for demonstration implementation of the recommendations presented herein, the WG members hope this work will inspire others in the community to apply these recommendations within their own research environments and share those experiences with the broader community. This joint report of the MaRDA WGs has focused on FAIR Data in two critical areas for materials researchers: (1) materials microscopy metadata and (2) laboratory information management. Future efforts could identify other critical materials research areas to concentrate FAIR data implementation. Funding requirements of the NSF FAIROS program limited participation in the materials microscopy and the LIMS WGs to the materials research community within the United States. To encourage international agreement, future endeavors in the global materials research community could target jointly designed and executed short-term projects funded by the national funding agencies of each participant. Global ventures comprised of mutual obligations and rewards could allow for stronger generalizations and insights leading to concrete actions to rapidly multiply successful implementations of FAIR data-driven AI across materials research communities and cognate science disciplines.

Author contributions

P.V., C.B., and L.B. initiated and organized the working groups. E.B., M.C., and M.T. co-chaired the materials microscopy metadata working group. J.T. and E.S. co-chaired the LIMS working group. J.T., E.B., and L.B. contributed the majority of the manuscript's text and J.T. served as corresponding author.

Funding

US National Science Foundation Findable Accessible Interoperable Reusable Open Science Research Coordination Networks (FAIROS RCN) NSF 22-553https://www.nsf.gov/pubs/2022/nsf22553/nsf22553.htmsupported this portion (NSF FAIROS RCN: 2226417) of the Materials Research Coordination Network as part of NSF's RCN program to advance and coordinate findable, accessible, interoperable, and reusable (FAIR) data.

Conflict of interest

On behalf of all authors, the corresponding author states that there is no conflict of interest.

Supplementary information

The online version contains supplementary material available at https://doi.org/10.1557/s43577-025-00882-2.

References

1. M. Scheffler, M. Aeschlimann, M. Albrecht, T. Bereau, H.-J. Bungartz, C. Felser, M. Greiner, A. Groß, C.T. Koch, K. Kremer, W.E. Nagel, M. Scheidgen, C. Wöll, C. Draxl, *Nature* **604**(7907), 635 (2022). https://doi.org/10.1038/s41586-022-04501-x
2. ...L.M. Ghiringhelli, C. Baldauf, T. Bereau, S. Brockhauser, C. Carbogno, J. Chamanara, S. Cozzini, S. Curtarolo, C. Draxl, S. Dwaraknath, A. Fekete, J. Kermode, C.T. Koch, M. Kühbach, A.N. Ladines, P. Lambrix, M.-O. Himmer, S.V. Levchenko, M. Oliveira, A. Michalchuk, R.E. Miller, B. Onat, P. Pavone, G. Pizzi, B. Regler, G.-M. Rignanese, J. Schaarschmidt, M. Scheidgen, A. Schneidewind, T. Sheveleva, C. Su, D. Usvyat, O. Valsson, C. Wöll, M. Scheffler, *Sci. Data* **10**(1), 626 (2023). https://doi.org/10.1038/s41597-023-02501-8
3. L.C. Brinson, L.M. Bartolo, B. Blaiszik, D. Elbert, I. Foster, A. Strachan, P.W. Voorhees, *MRS Bull.* **49**(1), 12 (2024). https://doi.org/10.1557/s43577-023-00498-4
4. L.J. Falling, *ACS Phys. Chem. Au* **4**(5), 420 (2024). https://doi.org/10.1021/acsphyschemau.4c00009
5. MaRDA: Materials Research Data Alliance (n.d.). https://marda-alliance.org

6. M.D. Wilkinson, M. Dumontier, I.J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L.B. Silva Santos, P.E. Bourne, J. Bouwman, A.J. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C.T. Evelo, R. Finkers, A. Gonzalez-Beltran, A.J.G. Gray, P. Groth, C. Goble, J.S. Grethe, J. Heringa, P.A.C. Hoen, R. Hooft, T. Kuhn, R. Kok, J. Kok, S.J. Lusher, M.E. Martone, A. Mons, A.L. Packer, B. Persson, P. Rocca-Serra, M. Roos, R. Schaik, S.-A. Sansone, E. Schultes, T. Sengstag, T. Slater, G. Strawn, M.A. Swertz, M. Thompson, J. Lei, E. Mulligen, J. Velterop, A. Waagmeester, P. Wittenburg, K. Wolstencroft, J. Zhao, B. Mons, *Sci. Data* **3**, 160018 (2016) https://doi.org/10.1038/sdata.2016.18 . arXiv: 1708.02002

7. A.A. Sinaci, F.J. Núñez-Benjumea, M. Gencturk, M.-L. Jauer, T. Deserno, C. Chronaki, G. Cangioli, C. Cavero-Barca, J.M. Rodríguez-Pérez, M.M. Pérez-Pérez, G.B. Laleci Erturkmen, T. Hernández-Pérez, E. Méndez-Rodríguez, C.L. Parra-Calderón, *Methods Inf. Med.* **59**(Suppl. 01), e21 (2020). https://doi.org/10.1055/s-0040-1713684

8. T.J. Jacobsson, A. Hultqvist, A. García-Fernández, A. Anand, A. Al-Ashouri, A. Hagfeldt, A. Crovetto, A. Abate, A.G. Ricciardulli, A. Vijayan, A. Kulkarni, A.Y. Anderson, B.P. Darwich, B. Yang, B.L. Coles, C.A.R. Perini, C. Rehermann, D. Ramirez, D. Fairen-Jimenez, D. Di Girolamo, D. Jia, E. Avila, E.J. Juarez-Perez, F. Baumann, F. Mathies, G.S.A. González, G. Boschloo, G. Nasti, G. Paramasivam, G. Martínez-Denegri, H. Näsström, H. Michaels, H. Köbler, H. Wu, I. Benesperi, M.I. Dar, I. Bayrak Pehlivan, I.E. Gould, J.N. Vagott, J. Dagar, J. Kettle, J. Yang, J. Li, J.A. Smith, J. Pascual, J.J. Jerónimo-Rendón, J.F. Montoya, J.-P. Correa-Baena, J. Qiu, J. Wang, K. Sveinbjörnsson, K. Hirselandt, K. Dey, K. Frohna, L. Mathies, L.A. Castriotta, M.H. Aldamasy, M. Vasquez-Montoya, M.A. Ruiz-Preciado, M.A. Flatken, M.V. Khenkin, M. Grischek, M. Kedia, M. Saliba, M. Anaya, M. Veldhoen, N. Arora, O. Shargaieva, O. Maus, O.S. Game, O. Yudilevich, P. Fassl, Q. Zhou, R. Betancur, R. Munir, R. Patidar, S.D. Stranks, S. Alam, S. Kar, T. Unold, T. Abzieher, T. Edvinsson, T.W. David, U.W. Paetzold, W. Zia, W. Fu, W. Zuo, V.R.F. Schröder, W. Tress, X. Zhang, Y.-H. Chiang, Z. Iqbal, Z. Xie, E. Unger, *Nat. Energy* **7**(1), 107 (2021). https://doi.org/10.1038/s41560-021-00941-3

9. N.T. Garabedian, P.J. Schreiber, N. Brandt, P. Zschumme, I.L. Blatter, A. Dollmann, C. Haug, D. Kümmel, Y. Li, F. Meyer, C.E. Morstein, J.S. Rau, M. Weber, J. Schneider, P. Gumbsch, M. Selzer, C. Greiner, *Sci. Data* **9**(1), 315 (2022). https://doi.org/10.1038/s41597-022-01429-9

10. J. Wise, A.G. De Barron, A. Splendiani, B. Balali-Mood, D. Vasant, E. Little, G. Mellino, I. Harrow, I. Smith, J. Taubert, K. Van Bochove, M. Romacker, P. Walgemoed, R.C. Jimenez, R. Winnenburg, T. Plasterer, V. Gupta, V. Hedley, *Drug Discov. Today* **24**(4), 933 (2019). https://doi.org/10.1016/j.drudis.2019.01.008

11. C. Cerchia, A. Lavecchia, *Drug Discov. Today* **28**(4), 103516 (2023). https://doi.org/10.1016/j.drudis.2023.103516

12. International Organization for Standardization, ISO/CASCO: ISO/IEC 17025:2017, *General Requirements for the Competence of Testing and Calibration Laboratories* (2018). https://www.iso.org/standard/66912.html

13. International Organization for Standardization, ISO/TC 176/SC 2: ISO 9001:2015, *Quality Management Systems - Requirements.* (2015). https://www.iso.org/standard/62085.html

14. E.S. Barnard, M.K.Y. Chan, E.A. Stach, J.A. Taillon, M.L. Taheri, J.W. Lau, L.M. Bartolo, L.C. Brinson, P.W. Voorhees, *MRS Bull.* **49**(3), 285 (2024). https://doi.org/10.1557/s43577-024-00676-y

15. B. Bayerlein, M. Schilling, M. Curran, C.E. Campbell, A.A. Dima, H. Birkholz, J.W. Lau, *Integ. Mater. Manuf. Innov.* **13**, 915 (2024). Accessed 18 Nov 2024. https://doi.org/10.1007/s40192-024-00378-y

16. I.G. Goldberg, C. Allan, J.-M. Burel, D. Creager, A. Falconi, H. Hochheiser, J. Johnston, J. Mellen, P.K. Sorger, J.R. Swedlow, *Genome Biol.* **6**(5), 47 (2005). https://doi.org/10.1186/gb-2005-6-5-r47

17. M. Könnecke, F.A. Akeroyd, H.J. Bernstein, A.S. Brewster, S.I. Campbell, B. Clausen, S. Cottrell, J.U. Hoffmann, P.R. Jemian, D. Männicke, R. Osborn, P.F. Peterson, T. Richter, J. Suzuki, B. Watts, E. Wintersberger, J. Wuttke, *J. Appl. Crystallogr.* **48**(1), 301 (2015). https://doi.org/10.1107/S1600576714027575

18. The NeXus Scientific Community: NXem - Nexus v2024.02 Documentation (2024). https://manual.nexusformat.org/classes/contributed_definitions/NXem.html

19. A.A. Guzman, V. Hofmann, O. Brendike-Mannix, *Electron Microscopy Glossary* (Helmholtz Metadata Collaboration, 2024). Accessed 20 Jan 2025. https://emglossary.helmholtz-metadaten.de

20. International Standard Organization, ISO/TC 202: ISO 5820:2024, *Microbeam Analysis – Hyper-dimensional Data File Specification (HMSA)* (2024). https://www.iso.org/standard/81733.html

21. S.L. Weibel, T. Koch, *Dlib. Mag.* **6**(12) (2000). https://doi.org/10.1045/december2000-weibel

22. K. Richards, R. White, N. Nicolson, R. Pyle, *A Beginner's Guide to Persistent Identifiers* (GBIF Secretariat, Copenhagen, 2011). https://doi.org/10.35035/mjgq-d052

23. J. Klump, K. Lehnert, D. Ulbricht, A. Devaraju, K. Elger, D. Fleischer, S. Ramdeen, L. Wyborn, *Data Sci. J.* **20**, 33 (2021). https://doi.org/10.5334/dsj-2021-033

24. Library of Congress, *Sustainability of Digital Formats: Planning for Library of Congress Collections,* TIFF, Revision 6.0 (2024). Accessed 14 Nov. 2024. https://www.loc.gov/preservation/digital/formats/fdd/fdd000022.shtml

25. M. Folk, G. Heber, Q. Koziol, E. Pourmal, D. Robinson, "An Overview of the HDF5 Technology Suite and Its Applications," in *AD'11: Proceedings of the EDBT/ICDT 2011 Workshop on Array Databases* (Associaton for Computing Machinery, New York, 2011), pp. 36–47. https://doi.org/10.1145/1966895.1966900

26. D.B. Newell, E. Tiesinga, *The International System of Units (SI),* 2019 edn. (Technical Report NIST SP 330-2019, National Institute of Standards and Technology, Gaithersburg, 2019). https://doi.org/10.6028/NIST.SP.330-2019; https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.330-2019.pdf

27. R. Hanisch, S. Chalk, R. Coulon, S. Cox, S. Emmerson, F.J. Flamenco Sandoval, A. Forbes, J. Frey, B. Hall, R. Hartshorn, P. Heus, S. Hodson, K. Hosaka, D. Hutzschenreuter, C.-S. Kang, S. Picard, R. White, *Nature* **605**(7909), 222 (2022). https://doi.org/10.1038/d41586-022-01233-w

28. M.M. Rathore, S.A. Shah, D. Shukla, E. Bentafat, S. Bakiras, *IEEE Access* **9**, 32030 (2021). https://doi.org/10.1109/ACCESS.2021.3060863

29. L.L. Haak, M. Fenner, L. Paglione, E. Pentz, H. Ratner, *Learn. Publ.* **25**(4), 259 (2012). https://doi.org/10.1087/20120404

30. P. Leach, M. Mealling, R. Salz, *A Universally Unique IDentifier (UUID) URN Namespace* (RFC Editor, July 2005). https://doi.org/10.17487/rfc4122; https://www.rfc-editor.org/info/rfc4122

31. M. Stocker, L. Darroch, R. Krahl, T. Habermann, A. Devaraju, U. Schwardmann, C. D'Onofrio, I. Häggström, *Data Sci. J.* **19**, 18 (2020). https://doi.org/10.5334/dsj-2020-018

32. R. Munroe, Standards (2011). Accessed 14 Nov 2024. https://xkcd.com/927

33. E. Prestat, F. de La Peña, J. Lähnemann, P. Jokubauskas, V. Tonaas Fauske, O. Pietsjoh, T. Ostasevicius, T. Nemoto, C. Francis, D.N. Johnstone, T. Furnival, N. Cautaerts, S. Somnath, pquinn-dls, J. Caron, K.E. MacArthur, M. Nord, P. Burdet, T. Aarholt, T. Poon, J.A. Taillon, N. Tappy, T. Slater, V. Migunov, DENSmerijn, M. Sarahan, "RosettaSciIO," *Zenodo* (2024). https://doi.org/10.5281/zenodo.8011666

34. A. Jacobsen, R. de Miranda Azevedo, N. Juty, D. Batista, S. Coles, R. Cornet, M. Courtot, M. Rosas, M. Dumontier, C.T. Evelo, C. Goble, G. Guizzardi, K.K. Hansen, A. Hasnain, K. Hettne, J. Heringa, R.W.W. Hooft, M. Imming, K.G. Jeffery, R. Kaliyaperumal, M.G. Kersloot, C.R. Kirkpatrick, T. Kuhn, I. Labastida, B. Magagna, P. McQuilton, N. Meyers, A. Montesanti, M. van Reisen, P. Rocca-Serra, R. Pergl, S.-A. Sansone, L.O.B. da Silva Santos, J. Schneider, G. Strawn, M. Thompson, A. Waagmeester, T. Weigel, M.D. Wilkinson, E.L. Willighagen, P. Wittenburg, M. Roos, B. Mons, E. Schultes, *Data Intell.* **2**, 10 (2020). https://doi.org/10.1162/dint_r_00024

35. G. Greene, J. Ragland, Z. Trautt, J. Lau, R. Plante, J. Taillon, A. Creuziger, C. Becker, J. Bennett, N. Blonder, L. Borsuk, C. Campbell, A. Friss, L. Hale, M. Halter, R. Hanisch, G. Hardin, L. Levine, S. Maragh, S. Miller, C. Muzny, M. Newrock, J. Perkins, A. Plant, B. Ravel, D. Ross, J.H. Scott, C. Szakal, A. Tona, P. Vallone, *A Roadmap for LIMS at NIST Material Measurement Laboratory* (NIST Technical Note 2216, National Institute of Standards and Technology, Gaithersburg, April 2022). https://doi.org/10.6028/NIST.TN.2216; https://nvlpubs.nist.gov/nistpubs/TechnicalNotes/NIST.TN.2216.pdf

36. P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, D. Kiela, Retrieval-augmented generation for knowledge-intensive NLP tasks (2020), Preprint, Version 4. arXiv. Version; https://doi.org/10.48550/ARXIV.2005.11401

37. S.G. Higgins, A.A. Nogiwa-Valdez, M.M. Stevens, *Nat. Protoc.* **17**(2), 17 (2022). https://doi.org/10.1038/s41596-021-00645-8.

38. F. Tristram, N. Jung, P. Hodapp, R.R. Schröder, C. Wöll, S. Bräse, *Adv. Funct. Mater.* **34**(20), 2303615 (2023) https://doi.org/10.1002/adfm.202303615

39. R.J. Hanisch, D.L. Kaiser, A. Yuan, A. Medina-Smith, B.C. Carroll, E.M. Campo, *NIST Research Data Framework (RDaF), Version 1.5* (NIST Special Publication NIST SP 1500-18r1, National Institute of Standards and Technology, Gaithersburg, May 2023). https://doi.org/10.6028/NIST.SP.1500-18r1; https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1500-18r1.pdf

40. R.V. Guha, D. Brickley, S. Macbeth, *Commun. ACM* **59**(2), 44 (2016). https://doi.org/10.1145/2844544

41. R. Albertoni, D. Browning, S.J.D. Cox, A. Gonzalez-Beltran, A. Perego, P. Winstanley, F. Maali, J. Erickson, Data Catalog Vocabulary (DCAT), Version 3 (2024). Accessed 13 Nov 2024. https://www.w3.org/TR/vocab-dcat-3

42. S. Leach-Murray, *Tech. Serv. Q.* **33**(1), 98 (2016). https://doi.org/10.1080/07317131.2015.1093855

43. G. Gueguen, E.L. Olson, N. Pfeiffer, OSF Metadata Application Profile (OSF MAP) (Open Science Framework, 2023). https://doi.org/10.17605/OSF.IO/8YCZR

44. B.J. Blaiszik, L. Ward, J. Gaff, A. Scourtas, B. Galewsky, Materials Data Facility Schemas (n.d.). Accessed 13 Nov. 2024. https://github.com/materials-data-facility/data-schemas

45. B. Blaiszik, A. Scourtas, K.J. Schmidt, ribhavb, E. Truelove, Z. Katok, A. Ambadkar, I. Darling, S. Wangen, N. Martinez, B. Cullen, L. Ward, C. Schneck, I. Foster, N. Pruyne, ryanchard, S.G. Baird, "MLMI2-CSSI/Foundry," *Zenodo* (2024). https://doi.org/10.5281/zenodo.10480757

46. J.A. Taillon, T.F. Bina, R.L. Plante, M.W. Newrock, G.R. Greene, J.W. Lau, *Microsc. Microanal.* **27**(3), 511 (2021). https://doi.org/10.1017/S1431927621000222

47. R.L. Plante, J.A. Taillon, J.W. Lau, G. Greene, M. Newrock, "Nexus-Experiment: An XML Schema for Describing Data Collected from Electron Microscopes" (National Institute of Standards and Technology, Gaithersburg, 2020). https://doi.org/10.18434/M32245; https://data.nist.gov/od/id/mds2-2245

48. K. Aur, *Sandia National Laboratories Ecosystem for Open Science: Metadata Schema v0.2 Description* (Technical Report SAND2020-12350 PE, Sandia National

Laboratories, September 2020). https://doi.org/10.2172/1777073; https://www.osti.gov/servlets/purl/1777073
49. M. Gould, "Hear us ROR! Announcing Our First Prototype and Next Steps," Version 1.0, *DataCite* (February 11, 2019). Accessed 13 Nov 2024. https://doi.org/10.5438/CYKZ-FH60; https://datacite.org/blog/hear-us-ror-announcing-our-first-prototype-and-next-steps
50. J.A. Kunze, E. Bermès, The ARK Identifier Scheme. Internet-Draft draft-kunze-ark-40, Internet Engineering Task Force (November 2024). Work in Progress. https://datatracker.ietf.org/doc/draft-kunze-ark/40
51. L. Lannom, B. Boesch, S. Sun, *Handle System Overview* (RFC Editor, November 2003). https://doi.org/10.17487/RFC3650; https://www.rfc-editor.org/info/rfc3650

Publisher's note