MaRDA FAIR Materials Microscopy and LIMS Data Working Groups' Community Recommendations [PREPRINT]

Joshua A. Taillon[1*], Edward S. Barnard[2], Laura M. Bartolo[3], Maria K. Y. Chan[4], Eric A. Stach[5], Mitra L. Taheri[6], L. Catherine Brinson[7], Peter W. Voorhees[8]

[1]Material Measurement Laboratory, National Institute of Standards and Technology, Boulder, CO, 80305, USA .
[2]Lawrence Berkeley National Laboratory, Berkeley, CA, 94720, USA .
[3]Center for Hierarchical Materials Design, Northwestern University, Evanston, IL, 60208, USA .
[4]Center for Nanoscale Materials, Argonne National Laboratory, Lemont, IL, 60439, USA .
[5]Department of Materials Science and Engineering, University of Pennsylvania, Philadelphia, PA, 19104, USA .
[6]Department of Materials Science and Engineering, Johns Hopkins University, Baltimore, MD, 21218, USA .
[7]Department of Mechanical Engineering and Materials Science, Duke University, Durham, NC, 27708, USA .
[8]Department of Materials Science and Engineering, Northwestern University, Evanston, IL, 60208, USA .

*Corresponding author(s). E-mail(s): joshua.taillon@nist.gov;
Contributing authors: esbarnard@lbl.gov;
laura.bartolo@northwestern.edu; mchan@anl.gov; stach@seas.upenn.edu;
mtaheri4@jhu.edu; cate.brinson@duke.edu;
p-voorhees@northwestern.edu;

Abstract

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 costs and benefits 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.

Keywords: infrastructure, microscopy data, laboratory information management systems, data/database, FAIR data

Introduction

Until recently, debate on 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 sub-fields of materials research, such as materials microscopy, as well as individuals' roles to plan, prepare and submit their research data in order to assemble 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 RCN) program to establish Research Coordination Networks in critical fields and geographical regions.[1] 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 artificial intelligence and machine learning. The MaRCN project involves six institutions: Johns Hopkins University (the lead institution), Duke University, Northwestern University, Purdue University, SUNY at Buffalo, 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

[1]U.S. 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.

for reusable health datasets [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 research and development in a biopharmaceutical private/public enterprise were implemented incorporating FAIR data principles in 2019 [10] and with more recent combination of artificial intelligence 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 over 100 attendees (primarily drawn from the United Stands and MaRDA membership) and invited presentations in two high priority areas given by leading experts in the fields:

- Mitra Taheri[2] (Johns Hopkins University) presented *Microscopy and FAIR Data*
- June Lau[3] (National Institute of Standards and Technology – NIST) presented *Electron microscopy facility data management: NexusLIMS (Laboratory Information Management Systems)*

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

[2]https://orcid.org/0000-0001-5349-1411
[3]https://orcid.org/0000-0002-5233-4956

datasets. 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-lifecycle 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 (WG) in these key, complementary areas important to the materials research community: 1) Materials Microscopy Metadata and 2) Laboratory Information Management Systems (LIMS). MaRDA Working Groups 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 Working Groups (WG) with Co-Chairs and Members comprised of recognized materials leaders and experts in the areas of Materials Microscopy and Laboratory Information Management Systems (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 Laboratory Information Management Systems into materials research data management workflows. Both groups concentrated their efforts on the types of "non-validated" environments typical in the experience of WG members, i.e. academic, government, and non-certified 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 Lab), and Mitra Taheri (Johns Hopkins University) with 10 members.[4] The LIMS Working Group was led by Co-Chairs Eric Stach (University of Pennsylvania) and Joshua Taillon (NIST) and 10 members[5] 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 Working Groups.[6]

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 MaRDA 2024 Annual Meeting (Virtual, 22 February 2024), the Midwest Microscopy and Microanalysis Meeting (Northwestern University, 15 March 2024) and the 2024 Spring MRS Meeting (Seattle, 22 April 2024) [14] and posted online [5].

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

[5]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 Corp).

[6]While 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.

These two MaRDA Working Groups 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 (ML) 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 below, 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 biological microscopy data [16]. It provides tools and resources to enable researchers to manage, share, and analyze large sets of (primarily biological) 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".

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 – HyperDimensional Spectral Data File Format: The HyperDimensional 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 hyperdimensional 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 ISO5820. HMSA datasets 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", "MeasurementMode", "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 Working Group 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 is well-documented, easily accessible, and valuable for future studies, including the future of AI-driven data analysis.

Comprehensive Metadata Capture

In general, more metadata is better. The complete capture of the context of a dataset should be represented in its metadata with an effort made to standardize naming and

organization of this information (see Fig. 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 4 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 this 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: *e.g.* 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, defocus. See Table 1 below.
4. Image data metadata should include technical information that describes the data and its 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].

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, 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). They 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 datasets, 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, instrument database, etc.) and unique identifiers and canonical persistent links are important in maintaining reliable connections between data. For an accessible introduction to PIDs, see Ref. 22.

Recording and standardizing data units in metadata is 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.

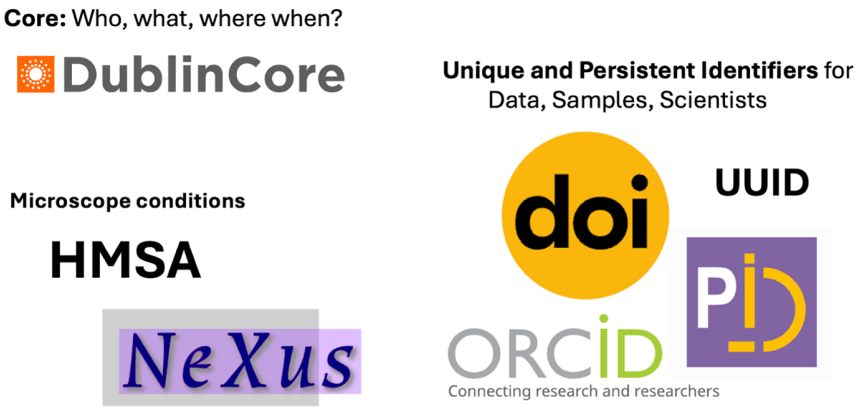

Fig. 1 Examples of existing metadata and identifier standards that can be used by materials EM community. For more information on each, please refer to the discussion in the text and consult the following references: Dublin Core [21], HMSA [20], NeXus and NXem [17, 18], DOI [22], ORCiD [29], UUID [30],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

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 Fig. 2). What is then also needed is efficient transformation of metadata standards from one to another. As mentioned before, ML techniques are beginning to make strides in defining ontologies and may soon provide automated translation between standards [15].

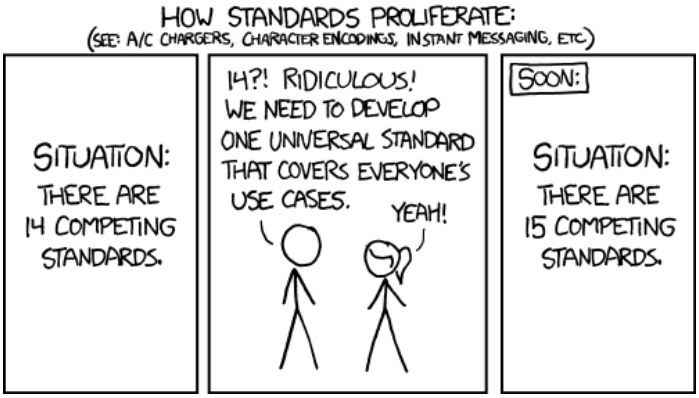

Fig. 2 A (humorous) perspective on how the desire to improve existing standards can lead to an endless proliferation of standards, making them all less effective in the process. [32]. We appreciate that materials microscopy metadata is susceptible to this same effect, and as such advocate for community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

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, 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, though 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 datasets [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 1.

Table 1 Examples of metadata (non-exhaustive) that should be included and what they should be called. Where possible, a formal vocabulary for terms (such as Ref. 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 & 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
...	...	...

Recommendations from the Laboratory Information Management Systems (LIMS) Working Group

The MaRDA LIMS Working Group (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. While simple data curation strategies (such as organizing data into folder hierarchies and embedding metadata in filenames) may 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 Working Groups) 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 Ref. 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. While 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 trade-off 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 Ref. 35.

Costs and Benefits of LIMS

While 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: *i.e.* "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., while 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 lab 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 $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 $30,000 to $150,000, depending on the need to procure storage hardware, decisions of onsite 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 on-going 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, e.g. 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 below 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 recommended above, 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 towards 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 Sec. 8)
- Experimental data is stored in, or system provides automatic conversion to, open data formats (making use of tools such as Ref. 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 above 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*[7] [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*. While we acknowledge this list of data models is non-exhaustive, 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 that 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, assistants, etc.)
  - What is the data (type, sample relationship, etc.)

[7] 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.

  - When/where was it collected (physical location and originating instrument)
  - Ideally, the system will allow for contextual information as well about why the data was 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 Section 8)

- **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, ARK ID, etc.)

  - this may 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, Processes, etc.

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 Fig. 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. While the 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 (Materials Research Data Alliance) Working Groups 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 Working Groups on Materials Microscopy and LIMS brought together stakeholders from universities, government labs, 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 Working Groups' accomplishments in materials data use and re-use lays a solid foundation to continue progress on building consensus for future FAIR data enhancements.

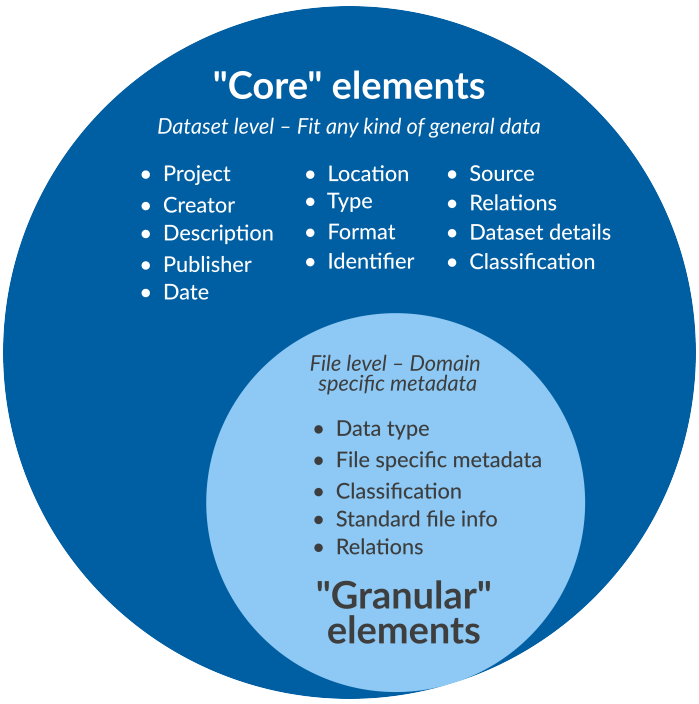

Fig. 3 An example of an extensible metadata structure split into "core" and "granular" elements from the *DCAT-eOS-AP* metadata model, adapted from Ref. 48.

In providing these recommendations to the materials research community, the Working Groups 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 may inspire individuals and organizations throughout our community to seriously consider incorporating formal microscopy metadata and LIMS approaches into their research data workflows. Furthermore, while 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 Working Groups has focused on FAIR Data in two critical areas for materials researchers: 1) materials microscopy metadata and 2) Laboratory Information Management. Future efforts might 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 Working Groups to the materials research community within the United States. To encourage international agreement, future endeavors in the global materials research community might 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 might 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.

Conflict of Interest Statement

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

References

[1] Scheffler, M., Aeschlimann, M., Albrecht, M., Bereau, T., Bungartz, H.-J., Felser, C., Greiner, M., Groß, A., Koch, C.T., Kremer, K., Nagel, W.E., Scheidgen, M., Wöll, C., Draxl, C.: FAIR data enabling new horizons for materials research. Nature **604**(7907), 635–642 (2022) https://doi.org/10.1038/s41586-022-04501-x

[2] Ghiringhelli, L.M., Baldauf, C., Bereau, T., Brockhauser, S., Carbogno, C., Chamanara, J., Cozzini, S., Curtarolo, S., Draxl, C., Dwaraknath, S., Fekete, A.,

Kermode, J., Koch, C.T., Kühbach, M., Ladines, A.N., Lambrix, P., Himmer, M.-O., Levchenko, S.V., Oliveira, M., Michalchuk, A., Miller, R.E., Onat, B., Pavone, P., Pizzi, G., Regler, B., Rignanese, G.-M., Schaarschmidt, J., Scheidgen, M., Schneidewind, A., Sheveleva, T., Su, C., Usvyat, D., Valsson, O., Wöll, C., Scheffler, M.: Shared metadata for data-centric materials science. Scientific Data **10**(1), 626 (2023) https://doi.org/10.1038/s41597-023-02501-8

[3] Brinson, L.C., Bartolo, L.M., Blaiszik, B., Elbert, D., Foster, I., Strachan, A., Voorhees, P.W.: Community action on FAIR data will fuel a revolution in materials research. MRS Bulletin **49**(1), 12–16 (2024) https://doi.org/10.1557/s43577-023-00498-4

[4] Falling, L.J.: A Vision for the Future of Materials Innovation and How to Fast-Track It with Services. ACS Physical Chemistry Au **4**(5), 420–429 (2024) https://doi.org/10.1021/acsphyschemau.4c00009

[5] MaRDA: Materials Research Data Alliance (MaRDA). https://marda-alliance.org

[6] Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., Silva Santos, L.B., Bourne, P.E., Bouwman, J., Brookes, A.J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C.T., Finkers, R., Gonzalez-Beltran, A., Gray, A.J.G., Groth, P., Goble, C., Grethe, J.S., Heringa, J., Hoen, P.A.C., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S.J., Martone, M.E., Mons, A., Packer, A.L., Persson, B., Rocca-Serra, P., Roos, M., Schaik, R., Sansone, S.-A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M.A., Thompson, M., Lei, J., Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J., Mons, B.: The FAIR Guiding Principles for scientific data management and stewardship. Scientific

Data **3**, 160018 (2016) https://doi.org/10.1038/sdata.2016.18 . arXiv: 1708.02002 ISBN: 20524463 (Electronic)

[7] Sinaci, A.A., Núñez-Benjumea, F.J., Gencturk, M., Jauer, M.-L., Deserno, T., Chronaki, C., Cangioli, G., Cavero-Barca, C., Rodríguez-Pérez, J.M., Pérez-Pérez, M.M., Laleci Erturkmen, G.B., Hernández-Pérez, T., Méndez-Rodríguez, E., Parra-Calderón, C.L.: From Raw Data to FAIR Data: The FAIRification Workflow for Health Research. Methods of Information in Medicine **59**(S 01), 21–32 (2020) https://doi.org/10.1055/s-0040-1713684

[8] Jacobsson, T.J., Hultqvist, A., García-Fernández, A., Anand, A., Al-Ashouri, A., Hagfeldt, A., Crovetto, A., Abate, A., Ricciardulli, A.G., Vijayan, A., Kulkarni, A., Anderson, A.Y., Darwich, B.P., Yang, B., Coles, B.L., Perini, C.A.R., Rehermann, C., Ramirez, D., Fairen-Jimenez, D., Di Girolamo, D., Jia, D., Avila, E., Juarez-Perez, E.J., Baumann, F., Mathies, F., González, G.S.A., Boschloo, G., Nasti, G., Paramasivam, G., Martínez-Denegri, G., Näsström, H., Michaels, H., Köbler, H., Wu, H., Benesperi, I., Dar, M.I., Bayrak Pehlivan, I., Gould, I.E., Vagott, J.N., Dagar, J., Kettle, J., Yang, J., Li, J., Smith, J.A., Pascual, J., Jerónimo-Rendón, J.J., Montoya, J.F., Correa-Baena, J.-P., Qiu, J., Wang, J., Sveinbjörnsson, K., Hirselandt, K., Dey, K., Frohna, K., Mathies, L., Castriotta, L.A., Aldamasy, M.H., Vasquez-Montoya, M., Ruiz-Preciado, M.A., Flatken, M.A., Khenkin, M.V., Grischek, M., Kedia, M., Saliba, M., Anaya, M., Veldhoen, M., Arora, N., Shargaieva, O., Maus, O., Game, O.S., Yudilevich, O., Fassl, P., Zhou, Q., Betancur, R., Munir, R., Patidar, R., Stranks, S.D., Alam, S., Kar, S., Unold, T., Abzieher, T., Edvinsson, T., David, T.W., Paetzold, U.W., Zia, W., Fu, W., Zuo, W., Schröder, V.R.F., Tress, W., Zhang, X., Chiang, Y.-H., Iqbal, Z., Xie, Z., Unger, E.: An open-access database and analysis tool for perovskite solar cells based on the FAIR data principles. Nature Energy **7**(1), 107–115 (2021)

https://doi.org/10.1038/s41560-021-00941-3

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

[10] Wise, J., De Barron, A.G., Splendiani, A., Balali-Mood, B., Vasant, D., Little, E., Mellino, G., Harrow, I., Smith, I., Taubert, J., Van Bochove, K., Romacker, M., Walgemoed, P., Jimenez, R.C., Winnenburg, R., Plasterer, T., Gupta, V., Hedley, V.: Implementation and relevance of FAIR data principles in biopharmaceutical R&D. Drug Discovery Today **24**(4), 933–938 (2019) https://doi.org/10.1016/j.drudis.2019.01.008

[11] Cerchia, C., Lavecchia, A.: New avenues in artificial-intelligence-assisted drug discovery. Drug Discovery Today **28**(4), 103516 (2023) https://doi.org/10.1016/j.drudis.2023.103516

[12] ISO/CASCO: ISO/IEC 17025:2017 General requirements for the competence of testing and calibration laboratories. International Organization for Standardization (2018). https://www.iso.org/standard/66912.html

[13] ISO/TC 176/SC 2: ISO 9001:2015 Quality management systems — Requirements. International Organization for Standardization (2015). https://www.iso.org/standard/62085.html

[14] Barnard, E.S., Chan, M.K.Y., Stach, E.A., Taillon, J.A., Taheri, M.L., Lau, J.W., Bartolo, L.M., Brinson, L.C., Voorhees, P.W.: NSF FAIROS Materials Research Data Alliance Working Groups to hold Town Hall Meeting at 2024 MRS Spring

Meeting & Exhibit. MRS Bulletin **49**(3), 285–286 (2024) https://doi.org/10.1557/s43577-024-00676-y . Publisher: Springer International Publishing

[15] Bayerlein, B., Schilling, M., Curran, M., Campbell, C.E., Dima, A.A., Birkholz, H., Lau, J.W.: Natural Language Processing-Driven Microscopy Ontology Development. Integrating Materials and Manufacturing Innovation (2024) https://doi.org/10.1007/s40192-024-00378-y . Accessed 2024-11-18

[16] Goldberg, I.G., Allan, C., Burel, J.-M., Creager, D., Falconi, A., Hochheiser, H., Johnston, J., Mellen, J., Sorger, P.K., Swedlow, J.R.: The Open Microscopy Environment (OME) Data Model and XML file: open tools for informatics and quantitative analysis in biological imaging. Genome Biology **6**(5), 47 (2005) https://doi.org/10.1186/gb-2005-6-5-r47

[17] Könnecke, M., Akeroyd, F.A., Bernstein, H.J., Brewster, A.S., Campbell, S.I., Clausen, B., Cottrell, S., Hoffmann, J.U., Jemian, P.R., Männicke, D., Osborn, R., Peterson, P.F., Richter, T., Suzuki, J., Watts, B., Wintersberger, E., Wuttke, J.: The NeXus data format. Journal of Applied Crystallography **48**(1), 301–305 (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] Guzman, A.A., Hofmann, V., Brendike-Mannix, O.: Electron Microscopy Glossary. Helmholtz Metadata Collaboration (2024). https://emglossary.helmholtz-metadaten.de/ Accessed 2025-01-20

[20] ISO/TC 202: ISO 5820:2024 Microbeam analysis — Hyper-dimensional data file specification (HMSA). International Organization for Standardization (2024).

https://www.iso.org/standard/81733.html

[21] Weibel, S.L., Koch, T.: The Dublin Core Metadata Initiative: Mission, Current Activities, and Future Directions. D-Lib Magazine **6**(12) (2000) https://doi.org/10.1045/december2000-weibel

[22] Richards, K., White, R., Nicolson, N., Pyle, R.: A Beginner's Guide to Persistent Identifiers. Technical report, GBIF (February 2011). https://doi.org/10.35035/mjgq-d052

[23] Klump, J., Lehnert, K., Ulbricht, D., Devaraju, A., Elger, K., Fleischer, D., Ramdeen, S., Wyborn, L.: Towards Globally Unique Identification of Physical Samples: Governance and Technical Implementation of the IGSN Global Sample Number. Data Science Journal **20**, 33 (2021) https://doi.org/10.5334/dsj-2021-033

[24] Library of Congress: TIFF, Revision 6.0. Publication Title: Sustainability of Digital Formats: Planning for Library of Congress Collections (2024). https://www.loc.gov/preservation/digital/formats/fdd/fdd000022.shtml Accessed 2024-11-14

[25] Folk, M., Heber, G., Koziol, Q., Pourmal, E., Robinson, D.: An overview of the HDF5 technology suite and its applications. In: Proceedings of the EDBT/ICDT 2011 Workshop on Array Databases, pp. 36–47. ACM, Uppsala Sweden (2011). https://doi.org/10.1145/1966895.1966900 . https://dl.acm.org/doi/10.1145/1966895.1966900

[26] Newell, D.B., Tiesinga, E.: The international system of units (SI): 2019 edition. Technical Report NIST SP 330-2019, National Institute of Standards and Technology, Gaithersburg, MD (August 2019). https://doi.org/10.6028/NIST.SP.330-2019 . https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.

330-2019.pdf

[27] Hanisch, R., Chalk, S., Coulon, R., Cox, S., Emmerson, S., Flamenco Sandoval, F.J., Forbes, A., Frey, J., Hall, B., Hartshorn, R., Heus, P., Hodson, S., Hosaka, K., Hutzschenreuter, D., Kang, C.-S., Picard, S., White, R.: Stop squandering data: make units of measurement machine-readable. Nature **605**(7909), 222–224 (2022) https://doi.org/10.1038/d41586-022-01233-w

[28] Rathore, M.M., Shah, S.A., Shukla, D., Bentafat, E., Bakiras, S.: The Role of AI, Machine Learning, and Big Data in Digital Twinning: A Systematic Literature Review, Challenges, and Opportunities. IEEE Access **9**, 32030–32052 (2021) https://doi.org/10.1109/ACCESS.2021.3060863

[29] Haak, L.L., Fenner, M., Paglione, L., Pentz, E., Ratner, H.: ORCID: a system to uniquely identify researchers. Learned Publishing **25**(4), 259–264 (2012) https://doi.org/10.1087/20120404

[30] Leach, P., Mealling, M., Salz, R.: A Universally Unique IDentifier (UUID) URN Namespace. Technical Report 1, The Internet Society (July 2005). https://doi.org/10.17487/rfc4122 . ISBN: 2013206534 Volume: 16. https://www.rfc-editor.org/info/rfc4122

[31] Stocker, M., Darroch, L., Krahl, R., Habermann, T., Devaraju, A., Schwardmann, U., D'Onofrio, C., Häggström, I.: Persistent Identification of Instruments. Data Science Journal **19**, 18 (2020) https://doi.org/10.5334/dsj-2020-018

[32] Munroe, R.: Standards (2011). https://xkcd.com/927/ Accessed 2024-11-14

[33] Prestat, E., De La Peña, F., Lähnemann, J., Jokubauskas, P., Tonaas Fauske, V., pietsjoh, Ostasevicius, T., Nemoto, T., Francis, C., Johnstone, D.N., Furnival, T., Cautaerts, N., Somnath, S., pquinn-dls, Caron, J., MacArthur, K.E., Nord, M.,

Burdet, P., Aarholt, T., Poon, T., Taillon, J.A., Tappy, N., Slater, T., Migunov, V., DENSmerijn, Sarahan, M.: RosettaSciIO. Zenodo (2024). https://doi.org/10.5281/zenodo.8011666

[34] Jacobsen, A., De Miranda Azevedo, R., Juty, N., Batista, D., Coles, S., Cornet, R., Courtot, M., Crosas, M., Dumontier, M., Evelo, C.T., Goble, C., Guizzardi, G., Hansen, K.K., Hasnain, A., Hettne, K., Heringa, J., Hooft, R.W.W., Imming, M., Jeffery, K.G., Kaliyaperumal, R., Kersloot, M.G., Kirkpatrick, C.R., Kuhn, T., Labastida, I., Magagna, B., McQuilton, P., Meyers, N., Montesanti, A., Van Reisen, M., Rocca-Serra, P., Pergl, R., Sansone, S.-A., Da Silva Santos, L.O.B., Schneider, J., Strawn, G., Thompson, M., Waagmeester, A., Weigel, T., Wilkinson, M.D., Willighagen, E.L., Wittenburg, P., Roos, M., Mons, B., Schultes, E.: FAIR Principles: Interpretations and Implementation Considerations. Data Intelligence **2**(1-2), 10–29 (2020) https://doi.org/10.1162/dint_r_00024

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

[36] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., Riedel, S., Kiela, D.: Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. arXiv. Version Number: 4 (2020).

https://doi.org/10.48550/ARXIV.2005.11401

[37] Higgins, S.G., Nogiwa-Valdez, A.A., Stevens, M.M.: Considerations for implementing electronic laboratory notebooks in an academic research environment. Nature Protocols **17**(2), 179–189 (2022) https://doi.org/10.1038/s41596-021-00645-8 . Publisher: Nature Research

[38] Tristram, F., Jung, N., Hodapp, P., Schröder, R.R., Wöll, C., Bräse, S.: The Impact of Digitalized Data Management on Materials Systems Workflows. Advanced Functional Materials, 2303615 (2023) https://doi.org/10.1002/adfm.202303615

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

[40] Guha, R.V., Brickley, D., Macbeth, S.: Schema.org: evolution of structured data on the web. Communications of the ACM **59**(2), 44–51 (2016) https://doi.org/10.1145/2844544

[41] Albertoni, R., Browning, D., Cox, S., Gonzalez-Beltran, A., Perego, A., Winstanley, P., Maali, F., Erickson, J.: Data Catalog Vocabulary (DCAT) - Version 3 (2024). https://www.w3.org/TR/vocab-dcat-3/ Accessed 2024-11-13

[42] Leach-Murray, S.: Figshare—Get credit for your research: Figshare.com. Technical Services Quarterly **33**(1), 98–99 (2016) https://doi.org/10.1080/07317131.2015.1093855

[43] Gueguen, G., Olson, E.L., Pfeiffer, N.: OSF Metadata Application Profile

(OSF MAP) (2023) https://doi.org/10.17605/OSF.IO/8YCZR . Publisher: Open Science Framework

[44] Blaiszik, B.J., Ward, L., Gaff, J., Scourtas, A., Galewsky, B.: Materials Data Facility Schemas. https://github.com/materials-data-facility/data-schemas Accessed 2024-11-13

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

[46] Taillon, J.A., Bina, T.F., Plante, R.L., Newrock, M.W., Greene, G.R., Lau, J.W.: NexusLIMS: A Laboratory Information Management System for Shared-Use Electron Microscopy Facilities. Microscopy and Microanalysis **27**(3), 511–527 (2021) https://doi.org/10.1017/S1431927621000222

[47] Plante, R.L., Taillon, J.A., Lau, J.W., Greene, G., Newrock, M.: NexusExperiment: an XML schema for describing data collected from electron microscopes. National Institute of Standards and Technology. Artwork Size: 2 files, 72.5 kB Pages: 2 files, 72.5 kB (2020). https://doi.org/10.18434/M32245 . https://data.nist.gov/od/id/mds2-2245

[48] Aur, K.: 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] Gould, M.: Hear us ROR! Announcing our first prototype and

next steps. Publisher: DataCite Version Number: 1.0 (2019). https://doi.org/10.5438/CYKZ-FH60 . https://datacite.org/blog/hear-us-ror-announcing-our-first-prototype-and-next-steps Accessed 2024-11-13

[50] Kunze, J.A., Bermès, E.: 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] Lannom, L., Boesch, L.C.B.P., Sun, S.: Handle System Overview. RFC Editor (2003). https://doi.org/10.17487/RFC3650 . https://www.rfc-editor.org/info/rfc3650

Authors Contributions

PV, CB, and LB initiated and organized the working groups. EB, MC, and MT co-chaired the materials microscopy metadata working group. JT and ES co-chaired the LIMS working group. JT, EB, and LB contributed the majority of the manuscript's text and JT served as corresponding author.

Funding

U.S. 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 supported 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.

Data availability

Not applicable.