1st Conference on Research Data Infrastructure

Harmonizing RDM

TIB-OP will set DOI with \TIBdoi

Published: 1970-01-01

FAIR research data with NOMAD

FAIRmat's distributed, schema-based research-data infrastructure to harmonize RDM in materials science

Markus Scheidgen[1] [https://orcid.org/0000-0002-8038-2277] , Sebastian Brückner[1,6] [https://orcid.org/0000-0002-5969-847X] , Sandor Brockhauser[1] [https://orcid.org/0000-0002-9700-4803] , Luca M. Ghiringhelli[1] [https://orcid.org/0000-0001-5099-3029] , Felix Dietrich[2] [https://orcid.org/0000-0002-2906-1769] , Ahmed E. Mansour[1] [https://orcid.org/0000-0002-3411-6808] , Martin Albrecht[6] [https://orcid.org/0000-0003-1835-052X] , Heiko B. Weber[7] [https://orcid.org/0000-0002-6403-9022] , Silvana Botti[34] [https://orcid.org/0000-0002-4920-2370] , Martin Aeschlimann[5] [https://orcid.org/0000-0003-3413-5029] , and Claudia Draxl[1] [https://orcid.org/0000-0003-3523-6657]

[1]Physics Department and IRIS Adlershof, Humboldt-Universität zu Berlin, Germany [2]School of CIT, Technical University of Munich, Munich, Germany [3]Research Center Future Energy Materials and Systems of the University Alliance Ruhr, Faculty of Physics and Astronomy, Ruhr University Bochum, Bochum, Germany [4]Institute for Solid State Theory and Optics, Friedrich Schiller University, Jena, Germany. [5]Department of Physics and Research Center OPTIMAS, University of Kaiserslautern, Kaiserslautern, Germany [6]Leibniz-Institut für Kristallzüchtung, Berlin, Germany [7]Department of Physics, Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany

Abstract: Scientific research is becoming increasingly data centric, which requires more effort to manage, share, and publish data. NOMAD is a web-based platform that provides research data management (RDM) for materials-science data. In addition to core RDM functions like uploading and sharing files, NOMAD automatically extracts structured data from supported file formats, normalizes, and converts data from these formats. NOMAD provides an extendable framework for managing not just files, but structured machine-actionable harmonized and inter-operable data. This is the basis for a faceted search with domain-specific filters, a comprehensive API, structured data entry via customizable ELNs, integrated data-analysis and machine-learning tools. NOMAD is run as a free public service and can additionally be operated by research institutes. Connecting NOMAD installations through the public services will allow a federated data infrastructure to share data between research institutes and further harmonize RDM within a large research domain such as materials science.

Keywords: materials science, research data management, electronic lab notebook, FAIR data, metadata

1 Introduction

In large research communities like materials science, researchers use many methods, instruments, tools, and workflows to produce large volumes of heterogeneous data files. The contained data describe semantically related research objects (like samples, materials, measurements) and it is believed that all combined data hold great potential for data re-use and artificial-intelligence-based solutions (therein including data mining and machine learning) [1], [2]. This is clearly being acknowledged not only by the research community but also by funding agencies, which are increasingly demanding coordinated efforts in research-data management (RDM) and the availability and longevity of open data by preserving and documenting all produced research data and metadata. This is the core mission of the German National Research Data Infrastructure (NFDI).

While individual researchers struggle with organizing and analyzing an increasing amount of data, communities face new challenges in making data Findable, Accessible, Interoperable, and Reproducible (FAIR) [3]. Collecting large amounts of heterogeneous files in generic repositories is not enough. Three key factors to FAIR data are (1) combining data with metadata and (2) putting all data into machine (and human) comprehensible and interoperable representations [4], [5], and (3) making the respective data-analysis tools accessible and let users directly execute them onto the stored data.

In this contribution, we present NOMAD, a distributed web-based platform for managing, sharing, and publishing FAIR data, based on rich domain-specific metadata and a homogeneous machine-comprehensible representation of data. NOMAD addresses RDM in two ways. First, it improves the data-centric workflows of individuals, laboratories, and research institutes by formalizing data acquisition, organizing and sharing data, homogenizing data for analysis, and integrating with analysis tools. This way, NOMAD provides the incentives and tools for research individuals to efficiently prepare FAIR (meta)data. Second, NOMAD allows researchers to collaborate on and publish harmonized data and serves communities as repository for FAIR data.

2 NOMAD software and service

NOMAD allows users to upload and manage files similar to other data repositories, e.g., Zenodo [6]. Beyond that, NOMAD also processes uploaded files and extracts machine-actionable data from over 60 formats and atomistic-computation codes and over 70 experimental techniques and applications by supporting the NeXus standard [7]. All processed data follow a schema that defines how the data are hierarchically structured, cross-referenced, and aggregated. This schema is called the *NOMAD Metainfo*. It defines a general domain-independent super structure, as well as highly detailed, specialized data from specific methods, tools, and programs. The NOMAD schema and processing can be extended by the community to support an ever increasing number of file types. NOMAD simply provides the framework and interfaces to harmonize data from heterogeneous sources. Furthermore, the NOMAD schema can be used to define workflows and Electronic Lab Notebooks (ELN) to augment data from files with structured data entry by human researchers; see also Figure 1.

NOMAD can run containerized tools such as Jupyter notebooks directly on data. This allows users to implement, run, and share their analysis tools alongside the analyzed data. The ability to run analysis tools in the browser, without installation, and specifically configured for a respective dataset lets researchers not just share data and analysis results, but the analysis itself and thereby drastically increases the reusability of data.

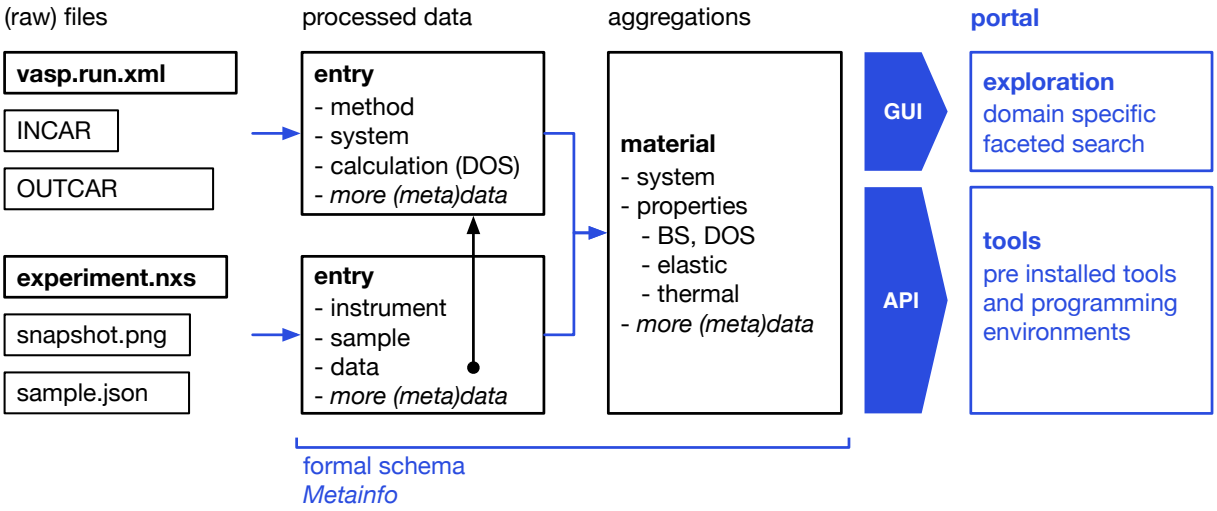

Figure 1. NOMAD provides a framework and interfaces for managing files alongside automatically extracted (meta)data.

NOMAD is build on open standards. For example, users can curate and publish datasets with a DOI and metadata based on DataCite [8], dataset metadata can be exported in various semantic web formats and is based on the DCAT vocabulary 2.0 [9]. NOMAD offers all computational results via the OPTIMADE [10] API specification. Other APIs are based on OpenAPI Specification [11]. Normalization of processed data if performed with widely accepted community software packages like MatID [12], ASE [13], or pymatgen [14].

The NOMAD software is used to operate a public and free NOMAD service that allows everyone to share and publish materials-science research data under the Create Commons Attribution License (cc-by) 4. This public NOMAD service contains as of April 2023 the data of over 12 million individual materials-science simulations and an increasing number of entries describing materials-science experiments. NOMAD is publicly available since 2014

and includes data from over 500 international authors. NOMAD includes the data of existing materials-science databases such as the Materials Project [15], AFLOW [16], OQMD [17], EELSDB [18], and the Perovskite Database Project [19], thereby overcoming heterogeneous database interfaces and providing all data with a singular API and in a common data representation.

The NOMAD software can also be independently operated by universities and other institutions to support local data management with independent data policies. Such self-managed installations are called *NOMAD Oases*, to distinguish them from the public NOMAD service. A NOMAD Oasis might be required when an institution needs to significantly customize the software for their needs, data volumes are too large to be conveniently transferred over the public internet, or for concerns about privacy or security. There will be the possibility to transfer data between different installations, and in order to adhere to the FAIR principles, the data (or at least metadata) in these Oases would ideally be made available to the public NOMAD service. NOMAD Oasis is used already by 15 of research institutes and can be used freely under the Apache 2 open-source license.

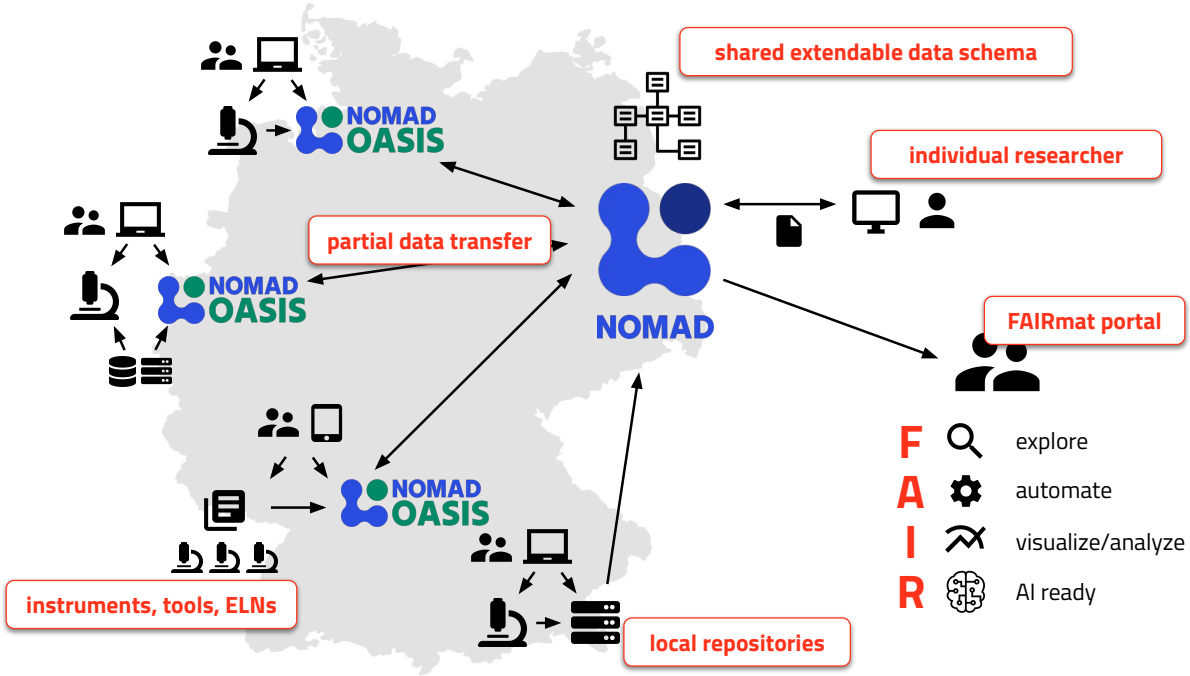

Figure 2. Distributed use of NOMAD in FAIRmat's federated data infrastructure.

The NFDI consortium FAIRmat develops NOMAD to build a federated FAIR-data infrastructure [1] (Figure 2). The NOMAD service will act as a central FAIRmat portal that will allow for accessing the data managed in NOMAD installations of participating institutes. NOMAD's processing and shared data schema provides harmonization and interoperability, while its distributed nature ensures findablity and accessibility across institutions. The potential to reuse and recontextualize data across a large heterogeneous community, should provide the incentives to further extend the schema, processing, and data analysis to foster continuous data harmonization.

Funding

NOMAD software development is funded by the the NFDI consortia FAIRmat (Deutsche Forschungsgemeinschaft, DFG, 460197019) and the NOMAD CoE (EU Horizon 2020, 951786); previous financial support was provided by the NOMAD CoE (EU Horizon 2020, 676580) and the Max-Planck Netzwerk BigMax. The Max Planck Computing and Data Facility (MPCDF) is hosting NOMAD's github and operating the public NOMAD service.

References

[1] M. Scheffler, M. Aeschlimann, M. Albrecht, *et al.*, "Fair data enabling new horizons for materials research," *Nature*, vol. 604, no. 7907, pp. 635–642, 2022. DOI: 10.1038/s41586-022-04501-x.

[2] L. Sbailò, Á. Fekete, L. M. Ghiringhelli, and M. Scheffler, "The nomad artificial-intelligence toolkit: Turning materials-science data into knowledge and understanding," *npj Computational Materials*, vol. 8, no. 1, p. 250, 2022. DOI: 10.1038/s41524-022-00935-z.

[3] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, *et al.*, "The fair guiding principles for scientific data management and stewardship," *Scientific data*, vol. 3, no. 1, pp. 1–9, 2016. DOI: 10.1038/sdata.2016.18.

[4] L. M. Ghiringhelli, C. Carbogno, S. Levchenko, *et al.*, "Towards efficient data exchange and sharing for big-data driven materials science: Metadata and data formats," *npj computational materials*, vol. 3, no. 1, p. 46, 2017. DOI: 10.1038/s41524-017-0048-5.

[5] L. M. Ghiringhelli, C. Baldauf, T. Bereau, *et al.*, "Shared metadata for data-centric materials science," *arXiv preprint arXiv:2205.14774*, 2022. DOI: 10.48550/arXiv.2205.14774.

[6] European Organization For Nuclear Research and OpenAIRE, *Zenodo*, en, 2013. DOI: 10.25495/7GXK-RD71. [Online]. Available: https://www.zenodo.org/.

[7] M. Könnecke, F. A. Akeroyd, H. J. Bernstein, *et al.*, "The nexus data format," *Journal of applied crystallography*, vol. 48, no. 1, pp. 301–305, 2015. DOI: 10.1107/S1600576714027575.

[8] J. Starr, J. Ashton, A. Barton, *et al.*, *Datacite metadata schema for the publication and citation of research data, version 3*, 2013. DOI: 10.5438/0008. [Online]. Available: https://schema.datacite.org/meta/kernel-3.0/doc/DataCite-MetadataKernel_v3.0.pdf.

[9] R. Albertoni, D. Browning, S. Cox, A. G. Beltran, A. Perego, and P. Winstanley, *Data catalog vocabulary (dcat) - version 2*, 2020. [Online]. Available: https://www.w3.org/TR/vocab-dcat-2/.

[10] C. W. Andersen, R. Armiento, E. Blokhin, *et al.*, "Optimade, an api for exchanging materials data," *Scientific data*, vol. 8, no. 1, pp. 1–10, 2021. DOI: 10.1038/s41597-021-00974-z.

[11] D. Miller, J. Whitlock, M. Gardiner, M. Ralphson, R. Ratovsky, and U. Sarid, *Openapi specification v3.0.3*, 2020. [Online]. Available: https://spec.openapis.org/oas/v3.0.3.

[12] L. Himanen, P. Rinke, and A. S. Foster, "Materials structure genealogy and high-throughput topological classification of surfaces and 2d materials," *npj Computational Materials*, vol. 4, no. 1, pp. 1–10, 2018. DOI: 10.1038/s41524-018-0107-6.

[13] A. H. Larsen, J. J. Mortensen, J. Blomqvist, *et al.*, "The atomic simulation environment—a python library for working with atoms," *Journal of Physics: Condensed Matter*, vol. 29, no. 27, p. 273 002, 2017. DOI: 10.1088/1361-648X/aa680e.

[14] S. P. Ong, W. D. Richards, A. Jain, *et al.*, "Python materials genomics (pymatgen): A robust, open-source python library for materials analysis," *Computational Materials Science*, vol. 68, pp. 314–319, 2013. DOI: 10.1016/j.commatsci.2012.10.028.

[15] A. Jain, S. P. Ong, G. Hautier, *et al.*, "Commentary: The materials project: A materials genome approach to accelerating materials innovation," *APL materials*, vol. 1, no. 1, p. 011 002, 2013. DOI: 10.1063/1.4812323.

[16] S. Curtarolo, W. Setyawan, G. L. Hart, *et al.*, "Aflow: An automatic framework for high-throughput materials discovery," *Computational Materials Science*, vol. 58, pp. 218–226, 2012. DOI: 10.1016/j.commatsci.2012.02.005.

[17] J. E. Saal, S. Kirklin, M. Aykol, B. Meredig, and C. Wolverton, "Materials design and discovery with high-throughput density functional theory: The open quantum materials database (oqmd)," *Jom*, vol. 65, no. 11, pp. 1501–1509, 2013. DOI: 10.1007/s11837-013-0755-4.

[18] P. Ewels, T. Sikora, V. Serin, C. P. Ewels, and L. Lajaunie, "A complete overhaul of the electron energy-loss spectroscopy and x-ray absorption spectroscopy database: Eelsdb.eu," *Microscopy and Microanalysis*, vol. 22, pp. 717–724, Feb. 2016, ISSN: 1435-8115. DOI: 10.1017/S1431927616000179. [Online]. Available: http://journals.cambridge.org/article_S1431927616000179.

[19] T. J. Jacobsson, A. Hultqvist, A. García-Fernández, *et al.*, "An open-access database and analysis tool for perovskite solar cells based on the fair data principles," *Nature Energy*, vol. 7, no. 1, pp. 107–115, Jan. 2022, ISSN: 2058-7546. DOI: 10.1038/s41560-021-00941-3. [Online]. Available: https://doi.org/10.1038/s41560-021-00941-3.