The Journal of Open Source Software

NOMAD: A distributed web-based platform for managing materials science research data

Markus Scheidgen [1*¶], **Lauri Himanen** [1*], **Alvin Noe Ladines** [1*], **David Sikter** [1*], **Mohammad Nakhaee** [1*], **Ádám Fekete** [1*], **Theodore Chang** [1*], **Amir Golparvar** [1*], **José A. Márquez** [1], **Sandor Brockhauser** [1], **Sebastian Brückner** [2], **Luca M. Ghiringhelli** [1], **Felix Dietrich** [3], **Daniel Lehmberg** [3], **Thea Denell** [1], **Andrea Albino** [1], **Hampus Näsström** [1], **Sherjeel Shabih** [1], **Florian Dobener** [1], **Markus Kühbach** [1], **Rubel Mozumder** [1], **Joseph F. Rudzinski** [1], **Nathan Daelman** [1], **José M. Pizarro** [1], **Martin Kuban** [1], **Cuauhtemoc Salazar** [1], **Pavel Ondračka** [4], **Hans-Joachim Bungartz** [3], **and Claudia Draxl** [1]

1 Department of Physics, Humboldt-Universität zu Berlin, Berlin, Germany **2** Leibniz Institut für Kristallzüchtung, Berlin, Germany **3** Department of Informatics, Technical University of Munich, Munich, Germany **4** Department of Plasma Physics and Technology, Masaryk University, Brno, Czech Republic ¶ Corresponding author * These authors contributed equally.

DOI: 10.21105/joss.05388

Software

- Review
- Repository
- Archive

Editor: Bonan Zhu

Reviewers:

- @arosen93
- @berquist
- @sgbaird

Submitted: 24 March 2023
Published: 26 September 2023

Summary

Materials science research is becoming increasingly data-driven, which requires more effort to manage, share, and publish data. NOMAD is a web-based application that provides data management for materials science research data. In addition to core data management functions like uploading and sharing files, NOMAD allows structured data entry using customizable forms providing the software with electronic laboratory notebook (ELN) functionalities. It automatically extracts rich metadata from supported file formats, normalizes and converts data from these formats, and provides a faceted search with materials science-specific filters based on extracted metadata. NOMAD integrates data analysis and machine learning tools. Installations of NOMAD can be connected to share data between research institutes and can publish data to an open central NOMAD service. The NOMAD software is distributed as a Docker image to create data management services and as a Python package to automate the client's use of these services.

Statement of need

In materials science, researchers use many methods, instruments, tools, and workflows to produce large volumes of heterogeneous data artifacts. The contained data often describes related research objects (materials, samples, or properties) and it is believed that all combined data hold great potential for data re-use and machine learning (Sbailò et al., 2022; Scheffler et al., 2022). This is clearly being acknowledged not only by the research community but also by funding agencies, which are increasingly demanding coordinated efforts in availability and longevity of open data by preserving and documenting all produced research data and meta-data.

While individual researchers struggle with organizing and analyzing more and more data artifacts, communities face new challenges in making data findable, accessible, inter-operable, and reproducible (FAIR) (Wilkinson et al., 2016). A key factor to FAIR data is to combine data

with meta-data and to put all data into machine and human comprehensible representations (Ghiringhelli et al., 2017, 2023).

Materials scientists require effective solutions for managing their research data, but they should not have to develop their own individual solutions. Hence, there is great demand in services (and software to run such services) that provide the mentioned features and make data FAIR. This is evident in the great number of published datasets on services like NOMAD (Draxl & Scheffler, 2018) (the main deployment of the NOMAD software), and an increasing number of materials science databases that all (re-)implement very similar functionality to publish their data.

NOMAD addresses these needs in two ways. First, NOMAD improves the data-driven workflows of individuals and small labs by formalizing data acquisition, organizing and sharing data, homogenizing and normalizing data for analysis, and integrating with analysis tools. This way, NOMAD provides the incentives and tools for research individuals to put the necessary efforts into preparing FAIR (meta-)data. Secondly, NOMAD allows to share or publish prepared data and can be used by communities as a repository for FAIR data.

Usage of NOMAD and related software

The NOMAD software is used to operate a public and free NOMAD service that allows everyone to share and publish materials science research data (https://nomad-lab.eu). This public NOMAD service contains over 12 million individual materials science simulations and around 50 thousand entries describing materials experiments or synthesis. NOMAD is publicly available since 2014 and includes data from over 500 international authors.

The NOMAD software can also be independently hosted by universities and other institutions when the use of the central service is not possible. 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 a specific need, the data volumes are too large to be conveniently transferred over the public internet, or when there are concerns about privacy or security. It should be noted that there is the possibility to transfer data between different installations, and in order to adhere to the FAIR principles, the data (or at least meta-data) in these Oases would ideally be transferred to the public NOMAD service. NOMAD Oasis is used by an increasing number of research institutes. NOMAD Oasis can be used freely as per our OSI license following the instruction in the NOMAD documentation.

The NFDI consortium FAIRmat uses NOMAD software as the bases for its federated FAIR data infrastructure (Scheffler et al., 2022).

OPTIMADE (Andersen et al., 2021) is an API specification (with associated software implementation) for materials science databases. NOMAD provides an implementation of the OPTIMADE specification and is an active part of the OPTIMADE consortium.

Other materials science databases (and the respective software) focus on publishing data that were produced with a specific framework and carefully curated by the group behind the database. Typical examples are databases of high-throughput simulations that try to systematically explore theoretical materials. Three of the larger databases of this kind are the Materials Project (Jain et al., 2013), AFLOW (Curtarolo et al., 2012), and OQMD (Saal et al., 2013). The raw data of these databases have also been published on NOMAD. The project AiiDA (Huber et al., 2020) allows scientists to design and run simulation workflows. AiiDA data can be published to AiiDA's materialscloud. There are also examples for experimental materials science databases, e.g. HTEM (Zakutayev et al., 2018).

NOMAD relies on many open source packages; a few more notable ones from the materials science domains are: *MatID*, a software package to identify material structure system types and symmetries (Himanen et al., 2018), *ASE*, a software package to manipulate material structures

in Python (Larsen et al., 2017), pymatgen, open-source python library for materials analysis (Ong et al., 2013), and *NeXus*, a file-format standard, schemas, and tools for experimental materials science data (Könnecke et al., 2015).

Acknowledgements

NOMAD software development is funded by the the German National Research Data Infrastructure (NFDI) consortium 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 Network BigMax. The Max Planck Computing and Data Facility (MPCDF) is hosting NOMAD's github and operating the public NOMAD service.

References

Andersen, C. W., Armiento, R., Blokhin, E., Conduit, G. J., Dwaraknath, S., Evans, M. L., Fekete, Á., Gopakumar, A., Gražulis, S., Merkys, A., & others. (2021). OPTIMADE, an API for exchanging materials data. *Scientific Data*, *8*(1), 1–10. https://doi.org/10.1038/s41597-021-00974-z

Curtarolo, S., Setyawan, W., Hart, G. L., Jahnatek, M., Chepulskii, R. V., Taylor, R. H., Wang, S., Xue, J., Yang, K., Levy, O., & others. (2012). AFLOW: An automatic framework for high-throughput materials discovery. *Computational Materials Science*, *58*, 218–226. https://doi.org/10.1016/j.commatsci.2012.02.005

Draxl, C., & Scheffler, M. (2018). NOMAD: The FAIR concept for big data-driven materials science. *Mrs Bulletin*, *43*(9), 676–682. https://doi.org/10.48550/arXiv.1805.05039

Ghiringhelli, L. M., Baldauf, C., Bereau, T., Brockhauser, S., Carbogno, C., Chamanara, J., Cozzini, S., Curtarolo, S., Draxl, C., Dwaraknath, S., & others. (2023). Shared metadata for data-centric materials science. *Scientific Data*, *10*(1), 626. https://doi.org/10.1038/s41597-023-02501-8

Ghiringhelli, L. M., Carbogno, C., Levchenko, S., Mohamed, F., Huhs, G., L"uders, M., Oliveira, M., & Scheffler, M. (2017). Towards efficient data exchange and sharing for big-data driven materials science: Metadata and data formats. *Npj Computational Materials*, *3*(1), 46. https://doi.org/10.1038/s41524-017-0048-5

Himanen, L., Rinke, P., & Foster, A. S. (2018). Materials structure genealogy and high-throughput topological classification of surfaces and 2D materials. *Npj Computational Materials*, *4*(1), 1–10. https://doi.org/10.1038/s41524-018-0107-6

Huber, S. P., Zoupanos, S., Uhrin, M., Talirz, L., Kahle, L., Häuselmann, R., Gresch, D., Müller, T., Yakutovich, A. V., Andersen, C. W., & others. (2020). AiiDA 1.0, a scalable computational infrastructure for automated reproducible workflows and data provenance. *Scientific Data*, *7*(1), 1–18. https://doi.org/10.1038/s41597-020-00638-4

Jain, A., Ong, S. P., Hautier, G., Chen, W., Richards, W. D., Dacek, S., Cholia, S., Gunter, D., Skinner, D., Ceder, G., & others. (2013). Commentary: The materials project: A materials genome approach to accelerating materials innovation. *APL Materials*, *1*(1), 011002. https://doi.org/10.1063/1.4812323

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., & others. (2015). The NeXus data format. *Journal of Applied Crystallography*, *48*(1), 301–305. https://doi.org/10.1107/S1600576714027575

Larsen, A. H., Mortensen, J. J., Blomqvist, J., Castelli, I. E., Christensen, R., Dułak, M., Friis, J., Groves, M. N., Hammer, B., Hargus, C., & others. (2017). The atomic simulation environment—a python library for working with atoms. *Journal of Physics: Condensed Matter*, *29*(27), 273002. https://doi.org/10.1088/1361-648X/aa680e

Ong, S. P., Richards, W. D., Jain, A., Hautier, G., Kocher, M., Cholia, S., Gunter, D., Chevrier, V. L., Persson, K. A., & Ceder, G. (2013). Python materials genomics (pymatgen): A robust, open-source python library for materials analysis. *Computational Materials Science*, *68*, 314–319. https://doi.org/10.1016/j.commatsci.2012.10.028

Saal, J. E., Kirklin, S., Aykol, M., Meredig, B., & Wolverton, C. (2013). Materials design and discovery with high-throughput density functional theory: The open quantum materials database (OQMD). *Jom*, *65*(11), 1501–1509. https://doi.org/10.1007/s11837-013-0755-4

Sbailò, L., Fekete, Á., Ghiringhelli, L. M., & Scheffler, M. (2022). The NOMAD artificial-intelligence toolkit: Turning materials-science data into knowledge and understanding. *Npj Computational Materials*, *8*(1), 250. https://doi.org/10.1038/s41524-022-00935-z

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

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., Silva Santos, L. B. da, Bourne, P. E., & others. (2016). The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, *3*(1), 1–9. https://doi.org/10.1038/sdata.2016.18

Zakutayev, A., Wunder, N., Schwarting, M., Perkins, J. D., White, R., Munch, K., Tumas, W., & Phillips, C. (2018). An open experimental database for exploring inorganic materials. *Scientific Data*, *5*(1), 1–12. https://doi.org/10.1038/sdata.2018.53