1. Binding letter of intent as advance notification or non-binding letter of intent

☑	Binding letter of intent (required as advance notification for proposals in 2019)
☐	Non-binding letter of intent (anticipated submission in 2020)
☐	Non-binding letter of intent (anticipated submission in 2021)

2. Formal details

Planned name of the consortium

FAIR Data Infrastructure for Materials Science and Related Fields

Acronym of the planned consortium

FAIRmat

Applicant institution

FAIR Data Infrastructure for Physics, Chemistry, Materials Science, and Astronomy e.V. (FAIR-DI) in explicit partnership with Humboldt Universität zu Berlin (HUB)[1] and Fritz-Haber-Institut der Max-Planck-Gesellschaft (FHI)[2]

Chairperson of the FAIR-DI Board: Prof. Dr. Matthias Scheffler

President of the HUB, Prof. Dr. Sabine Kunst

Executive Director of the FHI, Professor Dr. Robert Schlögl

Spokesperson

Prof. Dr. Claudia Draxl

Humboldt-Universität zu Berlin

Physics Department and IRIS Adlershof

Zum Großen Windkanal 6, 12489 Berlin

Phone: +49 30 2093 66363

Email: fairmat@fairdi.eu and claudia.draxl@physik.hu-berlin.de

3. Objectives, work programme and research environment

Research area of the proposed consortium (according to the DFG classification system)

32 (Physics), 31 (Chemistry), 406 (Materials Science) and 409 (Computer Science)

[1] HUB is a founding member of FAIR-DI e.V. and will provide office space and infrastructure for the employees of the FAIRmat consortium.

[2] The FHI is a founding member of FAIR-DI e.V. and will handle the financial administration, employment contracts, etc. for the FAIRmat consortium.

Concise summary of the planned consortium's main objectives and task areas

FAIRmat will install a **FAIR** (findable, accessible, interoperable, re-useable)[3] data infrastructure for the wider area of **mat**erials science[4]. Materials science represents a broad range of communities that can be characterized by either different classes of materials (e.g. semi-conductors, metals and alloys, soft matter, etc.), by different techniques (e.g. ranging from crystal-growth and synthesis to experimental and theoretical characterization by a multitude of probes), or by functionality (exemplified in here by battery materials, optoelectronics, heterogeneous catalysis, etc.). As a consequence, the data produced by materials-science are enormously heterogeneous and diverse in terms of the *4V* of Big Data, that are *V*olume (the amount of data), *V*ariety (heterogeneity of form and meaning of data), *V*elocity (rate at which data may change or new data arrive), and *V*eracity (uncertainty of data quality). These aspects are of different importance in the areas mentioned above. To cope with all the diversity, a bottom-up approach that satisfies the needs of the different areas/sub-communities is a must to foster acceptance by the community and participation of a large number of individual researchers and laboratories. FAIRmat sets out to tackle this challenge by a user-driven approach to develop easy-to-use tools and an infrastructure towards FAIR data processing, storage, curation, sharing, and future use of materials data.

Scientific results obtained by a certain experimental technique for a specific sample of a selected material are only meaningful and worth keeping if all individual steps are fully documented. This concerns the characterization of the sample, the description of the apparatus, as well as measurement conditions and the measured quantity. Likewise, computed data are only meaningful when method, approximations, code and code version, and well as computational parameters are known. In essence, we need a systematic metadata infrastructure, also covering ontologies. To address all of these essential aspects, we have identified several areas (further broken down into tasks) that are sketched in the following:

Area A – ***Materials Synthesis*** – is dedicated to the full characterization of samples and the corresponding synthesis / growth processes. Without this information, reproducibility of materials with given quality / properties will be hampered. The specific tasks will consider various synthesis routes, i.e., from the gas, liquid, and solid phases.

Area B – ***Experimental Materials Science*** – covers the microscopic characterization of materials by a broad variety of measurement techniques. Each of them comes with specific challenges concerning processing, curation, and storage, owing to differences in volume, velocity, data formats, etc. Metadata, ontologies, and workflows are instrument specific, and

[3] D. Wilkinson et al., Sci. Data 3, 160018 (2016).

[4] The initial focus of FAIRmat is on condensed (hard and soft) matter physics and chemistry and the function of these materials. With respect to the latter, FAIRmat addresses, in particular, the data challenges of energy-related research, e.g. heterogeneous catalysis, batteries, and optoelectronics.

key to ensure interpretability and re-usage of the data. FAIRmat will exemplify its approach in a first phase by a representative selection of techniques, i.e. tasks on electron microscopy and spectroscopy, angle-resolved photoemission, core-level spectroscopy, optical spectroscopy, atom-probe tomography, and scanning-probe microscopy. Other techniques will be added as soon as scientists raise such needs.

Area C – ***Computational Materials Science*** – deals with numerical techniques to compute materials properties. Such techniques differ in the theoretical concepts / theories, approximations and numerical recipes to solve the underlying equations. The obtained data also often differ in terms of file formats and units, depending on the employed code. Our general approach to tackle this diversity relies on the concepts developed in the European Center of Excellence NOMAD[5,6] that has developed parsers for 40 different community codes as well as a common data infrastructure. While NOMAD is currently largely focused on *ab initio* calculations, the tasks of Area C concern extensions towards excitations and strongly correlated materials, as well as towards classical (particle-based) simulations and multi-scale modeling. More codes and methods will be added as soon as scientists raise such needs.

Area D – ***Functional Materials*** – will demonstrate how the tools developed in the above areas will benefit different scientific communities. The specific tasks of this area will cover use cases on battery materials, heterogeneous catalysis, optoelectronics, spintronics & magnetism, biological physics applications, and artificial intelligence.

Area E – ***Digital Infrastructure*** – will be a common brace to all areas mentioned above. One of its tasks will be the linkage of metadata and ontologies developed in these areas such to provide a unified platform that will also be linked to international initiatives. Other tasks will be dedicated to processing and decentral storage, and creating a network of data hubs at different locations. A central metadata server together with a graphical user interface, including a Materials Encyclopedia[7], will allow for searching, accessing, and inspecting data from all over Germany (and even worldwide). All developed FAIRmat tools, from processing to post-processing, lab books, etc. will be provided as *open access* to the community. Tight collaborations with HPC centers will ensure the embedding into the overall NFDI landscape.

Area F – ***User Support, Training & Outreach*** – will reflect our concept for how to engage with the community, to allow researchers to make use and handle the FAIRmat tools.

Area G – ***Embedding into the Overall NFDI*** – will address synergies with other consortia and the interaction with the NFDI Directorate.

[5] https://nomad-coe.eu/

[6] C. Draxl and M. Scheffler, The NOMAD Laboratory: From Data Sharing to Artificial Intelligence, J. Phys. Mater. 2, 036001 (2019).

[7] Based on the NOMAD Encyclopedia (https://encyclopedia.nomad-coe.eu) that currently contains only computed data

Brief description of the proposed use of existing infrastructures, tools and services that are essential in order to fulfil the planned consortium's objectives (max. 1 page)

FAIRmat researchers have ample experience in building up and running data infrastructures, as evidenced by the non-profit association FAIR-DI (https://fairdi.eu). The most prominent pillar of FAIR-DI is the NOMAD Laboratory (https://nomad-coe.eu). It has created the NOMAD metainfo,[8] a metadata scheme for computational materials science. NOMAD has already developed parsers and normalizers for 40 different community codes and created a common data infrastructure hosting so far more than 50 million calculations, representing the biggest database in computational materials science worldwide. As the general concepts and rules of NOMAD coincide with the FAIR principles, NOMAD is a successful GoFAIR Implementation Network.[9] The NOMAD data, services, and hardware are currently hosted by the Max-Planck Computing and Data Facility (MPCDF).

Ample expertise on metadata and interoperability exists also through the DFG project GeRDI (Generic Research Data Infrastructure, https://www.gerdi-project.eu/) where some of the main developers are active within the FAIRmat team. Overall, FAIRmat can count on access to and cooperation with Germany's major data and computing facilities (LRZ [Munich], JSC [Jülich], MPCDF [Garching], ZIH [Dresden], TIB [Hannover], and others).

Our long-standing experience of building and maintaining an extensive data infrastructure will inform and guide the developments in the other areas, well knowing that every experimental or theoretical technique needs own solutions in order to be accepted by the community. Supporting the individual researchers and the materials science community in a wider sense is the topmost goal of FAIRmat. Satisfying these needs is ensured by the involvement and support of leading research organizations like the Condensed Matter Section (SKM) of the German Physical Society, many universities and research institutions, large-scale research facilities and research networks.

[8] https://metainfo.nomad-coe.eu

[9] https://www.go-fair.org/implementation-networks/overview/nomad/

Interfaces to other proposed NFDI consortia

Synergies and potential overlap with other consortia have been discussed with DAPHNE (previously DAPHNE and NDATA separately), NFDI4Chem, NFDI4Ing, NFDI4MSE, NFDI4cat, and MaRDI.

With **DAPHNE**, we agreed on a pilot project to identify the needs and interfaces in terms of data structures, metadata, and data exchange between the two consortia; to synergistically develop e-lab books; and to organize a workshop on metadata in 2021, also involving the consortium **MaRDI**. Members of the FAIRmat team have been actively pushing forward joint efforts related to metadata already during the last years. After a first workshop on metadata in computational materials science in 2016, another workshop in July 2019[10] will bring together people from experimental and computational materials science.

We have close contacts with both chemistry consortia, **NFDI4Chem** and **NFDI4cat**. Several FAIRmat members are also active in one or both of them. Heterogeneous catalysis is a component of FAIRmat, represented by renowned, worldwide-leading scientists of this field. Specifically, the chemical physics of surfaces of materials, chemical reactions at surfaces, and multi-scale modeling of heterogeneous catalysis will be explored in collaboration with NFDI4cat. Homogeneous and bio-catalysis, in contrast, are fully covered by NFDI4cat. Data aspects of theoretical / computational chemistry will be treated together with NFDI4Chem.

We are also in contact with the engineering consortia, **NFDI4Ing** und **NFDI4MSE**. With NFDI4MSE, we will collaborate in terms of metadata and ontologies for establishing links between the time and length scales addressed in FAIRmat with the scales of applied research as addressed by NFDI4MSE. This collaboration is already starting, as exemplified by a joint project involving - among other institutions - KIT, TIB, FHI, and Fraunhofer-IWM in the BMBF-funded project *STREAM (Semantische Repräsentation, Vernetzung und Kuratierung von qualitätsgesicherten Materialdaten)*.

[10] https://th.fhi-berlin.mpg.de/meetings/meta2019/

4. Cross-cutting topics

Please identify cross-cutting topics that are relevant for your consortium and that need to be designed and developed by several or all NFDI consortia.

Metadata, ontologies, and workflows are key to any consortium of a data infrastructure. In materials science they are particularly critical and complex. Being very domain- and even instrument- / code-specific, we aim at coordinating our efforts with other consortia such to optimally embed our developments into the overall NFDI (Area G). Likewise, we will team up with neighboring consortia to jointly organize workshops and training events.

Please indicate which of these cross-cutting topics your consortium could contribute to and how

We will contribute our experience with the NOMAD metainfo and our preliminary developments for experimental data. We will share our expertise in developing data archives, search engines, and graphical user interfaces. The Materials Encyclopedia (Area E) and our use cases (Area D) will make data accessible and comprehensible for different communities, and hence beneficial for fields outside materials science (e.g. finding materials for medical devices or energy applications).