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 2020)
☐	Non-binding letter of intent (anticipated submission in 2021)

2 Formal details

- Planned name of the consortium

FAIR Data Infrastructure for Condensed-Matter Physics and the Chemical Physics of Solids

- Acronym of the planned consortium

FAIRmat

- Applicant institution

Humboldt-Universität zu Berlin (HUB)

President of the HUB: Prof. Dr. Sabine Kunst

- Spokesperson

Claudia Draxl

Humboldt-Universität zu Berlin

Physics Department and IRIS Adlershof

FAIR-DI e.V.

Zum Großen Windkanal 6, 12489 Berlin

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

Phone: +49 30 2093 66363

- Co-applicant institutions and Co-spokespersons

Fritz Haber Institute of the Max Planck Society (FHI)

Faradayweg 4-6, 14195 Berlin

Managing director: Prof. Dr. Gerard Meijer

Co-Spokesperson: Matthias Scheffler, scheffler@fhi.mpg.de

Leibniz Institute for Crystal Growth (IKZ)
Max-Born-Str. 2, 12489 Berlin
Director: Prof. Dr. Thomas Schröder
Co-Spokesperson: Martin Albrecht, martin.albrecht@ikz-berlin.de

Max Planck Institute for Chemical Energy Conversion (MPI-CEC)
Stiftstrasse 34-36, 45470 Mülheim an der Ruhr
Managing Director: Prof. Dr. Robert Schlögl
Co-Spokesperson: Mark Greiner, mark.greiner@cec.mpg.de

Technical University of Munich (TUM)
Arcisstraße 21, 80333 Munich
President: Prof. Dr. Thomas F. Hofmann
Co-Spokesperson: Hans-Joachim Bungartz, bungartz@tum.de

Karlsruhe Institute of Technology (KIT)
Kaiserstraße 12, 76131 Karlsruhe
President: Prof. Dr.-Ing. Holger Hanselka
Co-Spokesperson: Christof Wöll, christof.woell@kit.edu

FAIR-DI e.V.
Zum Großen Windkanal 6, 12489 Berlin
Chairperson, Prof. Dr. Matthias Scheffler
Co-Spokespersons: Matthias Scheffler, **Christof Wöll**, **Claudia Draxl**
claudia.draxl@physik.hu-berlin.de, scheffler@fhi.mpg.de, christof.woell@kit.edu

- Participants

Martin Aeschlimann	Technical University of Kaiserslautern
Sören Auer	Library of Science and Technology (TIB)
Carsten Baldauf	Fritz Haber Institute of the Max Planck Society
Tristan Bereau	Max Planck Institute for Polymer Research, University of Amsterdam
Stefan Blügel	Research Center Jülich
Silvana Botti	Friedrich Schiller University Jena

Christoph J. Brabec	Friedrich-Alexander University Erlangen-Nürnberg
Malte Dreyer	Humboldt-Universität zu Berlin
Natascha Dropka	Leibniz Institute for Crystal Growth
Ralph Ernstorfer	Fritz Haber Institute of the Max Planck Society
Norbert Esser	Leibniz Institute for Analytical Sciences (isas)
Claudia Felser	Max Planck Institute for Chemical Physics of Solids
Luca Ghiringhelli	Fritz Haber Institute of the Max Planck Society
Roger Gläser	Leipzig University
Axel Groß	Ulm University
Marius Grundmann	Leipzig University
Aleksander Gurlo	TU Berlin
Thomas Hammerschmidt	Ruhr-University Bochum
Tamas Haraszti	University Aachen
Kerstin Helbig	Humboldt-Universität zu Berlin
Stefan Hecht	Leibniz Institute for Interactive Materials (DWI)
Thomas Heine	Technical University Dresden
Jürgen Janek	Justus Liebig University Giessen
Heinz Junkes	Fritz Haber Institute of the Max Planck Society
Josef A. Käs	Leipzig University
Wolfgang Klesse	Institute for Innovative Microelectronics (IHP)
Christoph Koch	Humboldt-Universität zu Berlin
Sarah Köster	Georg-August-University Göttingen
Kurt Kremer	Max Planck Institute for Polymer Research
Michael Krieger	Friedrich-Alexander University Erlangen-Nürnberg
Markus Kühbach	Max Planck Institute for Iron Research
Hermann Lederer	Max Planck Computing and Data Facility
Ingrid Mertig	Martin-Luther-University Halle-Wittenberg
Wolfgang Nagel	TU Dresden
Rossitza Pentcheva	University Duisburg-Essen

Dierk Raabe	Max Planck Institute for Iron Research
Alexander Reinefeld	Zuse Institute Berlin
Karsten Reuter	Fritz Haber Institute of the Max Planck Society
Raphael Ritz	Max Planck Computing and Data Facility
Cesar Rodriguez-Emmenegger	Leibniz Institute for Interactive Materials (DWI)
Erich Runge	ex officio Chairperson of the *Sektion Kondensierte Materie* of the DPG
Markus Scheidgen	Humboldt-Universität zu Berlin
Robert Schlögl	Fritz Haber Institute of the Max Planck Society
Thomas Schröder	Leibniz Institute for Crystal Growth
Godehard Sutmann	Jülich Supercomputing Centre
Annette Trunschke	Fritz Haber Institute of the Max Planck Society
Thomas Unold	Helmholtz Zentrum Berlin
Roser Valenti	Goethe University Frankfurt
Holger von Wenckstern	Leipzig University
Dieter Weber	Research Center Jülich
Heiko Weber	Friedrich-Alexander University Erlangen-Nürnberg
Christian Wenger	Institute for Innovative Microelectronics
Stefan Wesner	Ulm University
Joachim Wosnitza	Helmholtz Zentrum Dresden Rossendorf

3 Objectives, work programme and research environment

- Research area of the proposed consortium (according to the DFG classification system: https://www.dfg.de/download/pdf/dfg_im_profil/gremien/fachkollegien/ amtsperiode_2020_2023/fachsystematik_fachkollegien_amtsperiode_2020-2023.pdf)

307 Physik der kondensierten Materie

Specifically, FAIRmat attempts to cover the full comprehensiveness of the *Sektion Kondensierte Materie* of the German Physical Society (DPG) with its 12 *Fachverbände*. https://www.dpg-physik.de/vereinigungen/fachlich/skm/sektion-kondensierte-materie?set_language=en

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

FAIRmat will install a **FAIR**[1] data infrastructure for the wider area of ***condensed-matter physics*** and the ***chemical physics of solids***. This represents a very broad range of different communities that can be characterized by either different classes of condensed matter (e.g. semiconductors, metals and alloys, soft and biological 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 here by battery materials, optoelectronics, catalysts, etc.). As a consequence, the data produced by the community are enormously heterogeneous and diverse in terms of the 4V of Big Data, i.e. Volume (amount of data), Variety (heterogeneity of form and meaning of data), Velocity (rate at which data may change or new data arrive), and Veracity (uncertainty of data quality). Also note that many research data produced today may appear irrelevant in the context they have been produced. Being regarded as "waste", they are not published. However, they may turn out highly valuable for other purposes. So, the *R* in FAIR (reusability) also means "store, share, and recycle the waste!" To cope with all the diversity and complexity, a bottom-up approach that satisfies the needs of the different 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. For the latter, a major goal of FAIRmat is ***making data artificial-intelligence (AI) ready***.

Data obtained by a certain experimental technique for a specific sample of a selected material are only worth keeping if the sample is fully characterized and apparatus and measurement conditions as well as the measured quantity are described in detail. Likewise, computed data are only meaningful when method, approximations, code and code version, as well as all computational parameters are known. In essence, we need an extensive annotation, i.e. a systematic metadata catalogue, also covering ontologies. This also includes the description of provenance and data quality (their usefulness for a given context). To address all these aspects and to support basic science and individual researchers, we have identified several task areas (in short called Areas, further broken down into tasks) that are sketched in the following:

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

[1] FAIR = findable, accessible, interoperable, reusable: D. Wilkinson et al., Sci. Data 3, 160018 (2016).

Area B – ***Experiment*** – 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. FAIRmat will exemplify its approach in a first phase by a representative selection of experimental probes.

Area C – ***Theory and Computation*** – deals with numerical techniques to compute materials properties. Such techniques differ in the theoretical concepts, approximations, and numerical recipes, depending on the employed software. Our general approach to tackle this diversity relies on the concepts developed by the NOMAD Laboratory[2]. 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.

Area D – ***Digital Infrastructure*** – will be a common brace to all other areas. Specific tasks will be dedicated to processing and decentral storage, creating a network of data hubs at different locations. The FAIRmat *Portal*, a web-based graphical user interface, including a *Materials Encyclopedia*[3], will allow for searching, accessing, and inspecting (meta)data from all over Germany (and worldwide). FAIRmat will develop and provide tools, from processing to post-processing, including analysis by artificial intelligence, and it will provide guidance and advancements of electronic lab-books (ELN), laboratory information management systems (LIMS), etc. Tight collaborations with HPC centers will ensure the embedding into the overall NFDI landscape.

Area E – ***Use-Case Demonstrators*** – will highlight how the tools developed in the above areas will benefit different scientific communities and demonstrate hand-shakes and potential synergies with other consortia. The specific tasks of this area will cover in the first phase use cases on battery materials, heterogeneous catalysis, optoelectronics, spintronics & magnetism, metal-organic frameworks, biological physics applications, and artificial intelligence.

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. We will offer a variety of workshops and other training opportunities and provide support for connecting data hubs to the overall data infrastructure.

Area G – ***Administration and Coordination*** – will deal with all coordination and management issues and the embedding into the Overall NFDI. As such, it will address synergies with other consortia and the interaction with the NFDI Directorate.

---

[2] C. Draxl and M. Scheffler, The NOMAD Laboratory: From Data Sharing to Artificial Intelligence, J. Phys. Mater. 2, 036001 (2019); https://doi.org/10.1088/2515-7639/ab13bb
[3] Based on the NOMAD Encyclopedia (https://encyclopedia.nomad-coe.eu) currently containing 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

FAIRmat researchers have ample experience in building and running data infrastructures, evidenced by the non-profit association FAIR-DI e.V. Its most prominent pillar is the NOMAD Laboratory. It has created the NOMAD Metainfo, a metadata schema for computational materials science, and developed parsers / interpreters / normalizers for the 40 most important large-scale computer codes of the community. Hosting so far more than 100 million calculations (summer 2020), it represents the worldwide biggest data collection of its kind. Note that in the last year, the number has doubled. The NOMAD Repository & Archive are fully FAIR, even spearheading other "FAIR activities". NOMAD was among the first GoFAIR Implementation Networks. The European Center of Excellence *NOMAD* is a component of the NOMAD Laboratory, which is now focusing on software developments for upcoming exascale computers.

Ample expertise on metadata and interoperability exists also through the DFG project GeRDI (Generic Research Data Infrastructure). Overall, FAIRmat can count on access to and cooperation with Germany's major data and computing facilities (JSC [Jülich], MPCDF [Garching], ZIH [Dresden], ZIB [Berlin], kiz [Ulm], LRZ [Munich], TIB [Hanover], etc.). FAIRmat already created publications and organized workshops on metadata and ontology challenges of the condensed-matter and chemical-physics community.[4,5,6] FAIRmat's recent *Satellite Workshop on Data Acquisition in Angle-Resolved Photoemission Spectroscopy* was the last in the series. The big virtual FAIR-DI Conference on a *FAIR Data Infrastructure for Materials Genomics* brought the international community together and hosted the *Satellite Workshop on NFDI@Teaching*.

Supporting the individual researchers and the entire community 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 DPG, many universities and research institutions, large-scale research facilities and research networks. Most of the SKM divisions are already active within FAIRmat, others will be integrated as soon as they are ready to join. The installation of a FAIR infrastructure will only be successful if it involves also the international community. Memoranda of Understanding (MuO) have been signed with NIST (USA), Shanghai University and their Materials Genome Institute (China), and the Finish IT Center CSC. Other international partners are directly involved via their membership in FAIR-DI e.V.

---

[4] L.M. Ghiringhelli, et al, *Towards efficient data exchange and sharing for big-data driven materials science: metadata and data formats*. npj Computational Materials 3, 46 (2017); https://doi.org/10.1038/s41524-017-0048-5
[5] L.M. Ghiringhelli et al., *Shared Metadata and Data Formats for Big-Data-Driven Materials Science,* to be published.
[6] Shared metadata and data formats for Big-Data Driven Materials Science: A NOMAD-FAIRDI workshop, Berlin Adlershof, Germany, July 8-12, 2019; https://th.fhi-berlin.mpg.de/meetings/meta2019/

- Interfaces to other proposed NFDI consortia: brief description of existing agreements for collaboration and/or plans for future collaboration

Synergies and potential overlap with other consortia have been discussed with many consortia. 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 ELNs; and to organize a workshop on metadata in 2022, also involving the consortium **MaRDI**. We recall that members of the FAIRmat team have been actively pushing forward joint efforts related to metadata already during the last years. With MaRDI, we will also collaborate on the developments of metadata and ontology descriptions for data from “computer experiments” (Area C).

We have close contacts with both chemistry consortia**, NFDI4Chem** and **NFDI4cat**. Several FAIRmat members are also active in one or both. The condensed-matter and chemical physics of heterogeneous catalysis are 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, where the latter will extend these studies towards engineering. Data aspects of theoretical / computational chemistry as well as ELNs will be treated together with NFDI4Chem. There are common interests in quantum chemistry where FAIRmat already serves several important computer codes, where parsers, normalizers, converters, and metadata have been developed for Gaussian, GAMESS, MOLCAS, DMol3, NWChem, ORCA, Turbomole, etc. by FAIRmat’s component NOMAD already.

We are also in close contact with the engineering consortia, **NFDI4Ing** und **NFDI-MatWerk**. With the latter, we will collaborate in terms of metadata and ontologies for establishing links between the time and length scales addressed in FAIRmat with the larger scales of applied research as addressed by NFDI-MatWerk. Their strong engineering expertise on processing history and microstructure information is fully complementary to FAIRmat’s basic-science focus on electronic- and atomic-structure modelling. This enables a digital representation (ontology) of materials under realistic conditions. With NFDI4Ing we will collaborate on open-access tools, e.g. ELN solutions, data analysis methods, and workflows. Further, also note agreements with **PUNCH4NFDI**, **NFDI4Phys**, **DataPLANT**, **NFDI4MED**, **NFDI4MobilTech, NFDI4DataScience**, and **NFDIxCS**.

Boundaries between FAIRmat and other consortia as well as hand-shakes, and synergies will be highlighted by our *Use-Case Demonstrators* (Area E).

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.

Overall, an in-depth description of the cross-cutting topics concerning basically all consortia can be found in the *Leipzig-Berlin-Erklärung zu NFDI-Querschnittsthemen der Infrastrukturentwicklung*. Most of them are relevant for our consortium. Thus, we don't go in detail here.

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

Metadata, ontologies, and workflows are key to any consortium of a data infrastructure. In our research areas they are particularly critical and complex. Being very domain- and even instrument- or code-specific, we aim at coordinating our efforts with other consortia such to optimally embed our developments into the overall NFDI (Area G). More specifically, we will contribute our experience with the NOMAD Metainfo, our international contacts and collaborations, and our preliminary developments for experimental data. We will share our expertise in developing data archives, search engines, and graphical user interfaces. The *FAIRmat Portal* (Area D) and our use cases (Area E) will make data accessible and comprehensible for different communities, and hence beneficial for fields outside condensed-matter and chemical physics (e.g. finding materials for medical devices, energy, environment and mobility applications, and many more).

Furthermore, the systematic use of ELNs and LIMS is key to a successful NFDI. Here, we will make sure that our activities will be closely coordinated with all other consortia from physics, chemistry, and engineering. For instance, we will initiate workshops on ELNs and LIMS, also including stakeholders from other countries in the organization. This concerns in particular, people and institutions involved in FAIR-DI e.V. and those where we have signed MoUs with.

Another most important point concerns training and education. Interdisciplinary competencies, i.e., combining domain-specific skills with IT knowledge, are a crucial prerequisite for setting up a data infrastructure and key for its long-term goals and success. Moreover, we need to train the young generation making use of and exploiting research data for the benefit of our society. For all these reasons, we have been exploring possibilities to install new concepts in academic education. In this context, we are also in contact with the Young DPG (see also the *Satellite Workshop on NFDI@Teaching*). As a precursor to potential new curricula, we already offer courses on big-data driven science (e.g. at HU Berlin [summer semester 2020] and as video lectures within the Max Planck Graduate Center for Quantum Materials [winter semester 2020]).