**Please address the following aspects in your letter of intent**

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

*[Please indicate clearly whether your document is a binding letter of intent as advance notification or a non-binding letter of intent.]*

| | |
|---|---|
| ☐ | Binding letter of intent (required as advance notification for proposals in 2019) |
| **X** | 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
  NFDI for Interdisciplinary Research and Collaboration
- Acronym of the planned consortium
  InterdisciplinaryNFDI
- Applicant institution
  Universität Hamburg
  Mittelweg 177, 20148 Hamburg
  Prof. Dr. Dieter Lenzen
- Spokesperson
  Dr. Stefan Thiemann
  Universität Hamburg, Zentrum für nachhaltiges Forschungsdatenmanagement
  Monetastr. 4, 20146 Hamburg
  Phone: +49 40 42838 3844
  E-Mail: stefan.thiemann@uni-hamburg.de

## 3 Objectives, work programme and research environment

- Research area of the proposed consortium (according to the DFG classification system)
  www.dfg.de/download/pdf/dfg_im_profil/gremien/fachkollegien/amtsperiode_2016_2019/fachsystematik_2016-2019_en_grafik.pdf
  Humanities (11), Natural Sciences (31, 32), Materials Science and Engineering (43), Computer Science (44), Medicine (22)
- Concise summary of the planned consortium's main objectives and task areas

  NFDI has a more or less strict focus on research data from single scientific disciplines, or even research areas. This might be a good idea for large research areas such as astronomy, physics, or climate research. However, nowadays research is often organized in an interdisciplinary way, and heterogeneous data from different science areas need to be combined in such a way that systematic reuse of (parts of) data is ensured to the benefit of making progress in science.

  The described consortium will focus on data management, linking and interoperability issues for research data from different research areas. Especially in larger research projects, such as Research Training Groups, Forschungsgruppen, Sonderforschungsbereiche, or Clusters of Excellence, researchers from different areas intensively co-operate. They produce interdisciplinary research data of various kinds, and besides managing storage and organizing metadata descriptions, the interesting questions are: How can we link or connect heterogeneous data? Is it possible to do this automatically? Can we integrate data from other groups in the world? Can we organize efficient visualization especially for large data sets from the natural sciences and make interaction with more abstract data as easy as clicking on an image of an artefact? In order to provide answers to these questions we set up the following objectives for an NFDI consortium and analyse known needs.

  The key objective of the prospective consortium is to methodologically **organize data management for interdisciplinary research and collaboration**, and, as a

first step, we consider contributions from the humanities, natural sciences, and computer science as an example. We would like to emphasize that the problem of supporting interdisciplinary data management in NFDI is also relevant for single scientific disciplines that are quite large. Thus, similar examples can be given for large areas such as medical research (with interdisciplinary data as diverse as radiology data, omics data, relational electronic health records). As a second step of our consortium evolution, we will investigate an extension of the main ideas of InterdisciplinaryNFDI to medical data as well.

For achieving the key objective, different subgoals are defined. First, consortium activities have the goal of contributing to **solving the data integration problem** with a standard for **describing and managing compound data objects** defined w.r.t. base objects described using standards from respective disciplines. Compound objects are sometimes also called digital documents. For instance, Bayerische Staatsbibliothek (BSB) pursues an approach where base data such as pictures, materials science data as well as text data (in PDF or other formats) are grouped into compound objects, which are then associated with individual artefacts. To give another example, in the DFG Cluster of Excellence “Understanding Written Artefacts” (UWA) at Universität Hamburg, materials science data, such as x-ray data or spectrographic data of, e.g., clay tablets, are grouped with other data into compounds as well, or, just to mention another example, materials science data stemming from analysis of historical ink might be associated with written artefacts (physical objects) in a compound together with images and texts. While base data can have metadata (which are already standardized or might be standardized according to the evolution of the associated disciplines), compound data can have metadata as well.

Second, the consortium will investigate how data that is (automatically) **derived** from base data for certain **interpretation tasks** can be systematically attached to compound data (or hierarchies of compound data). In other words, the goal is to define data representations for introducing and **managing layers of derived data in compounds**, **defining various kinds of abstractions of data** used for solving different research problems. For instance, text data can be layered with (subjective and task-specific) symbolic descriptions for named entities (found in certain locations in texts) as well as relations between entities somehow mentioned in text data and subjectively considered relevant for answering certain research questions. While some approaches confuse derived data with metadata, the goal of the consortium is to advocate a clear separation of metadata, giving information about data production and scope, and derived data, describing abstractions (or interpretations) of data itself. Derived data can have metadata as well, and, in turn, can very well be compound data. Compounds can be hierarchically organized, and thus, we speak of **vertical data integration**. Besides thinking in terms of a single large data centre for some disciplines it is necessary to have competencies and infrastructure for research data management at most universities/research institutions. A task for the NFDI is building a network, and the consortium aims at providing something like a **federated dataset search** with automatic **(horizontal) data integration facilities involving declarative mapping rules** for connecting also to European and global scientific communities.

Third, anticipating recent and new technology of information retrieval, the consortium will investigate how task-specific **links between data can be described, (automatically) derived, and managed**. Note that links need not be explicit, such as in linked open data approaches, but can very well be defined implicitly and possibly in a task-specific way by exploiting named entity recognition techniques in texts (same entities mentioned in different texts give a link between the texts), pattern recognition in multidimensional vector data (similar patterns mentioned in different locations can also define links). Given an association defined for one modality (e.g., spectroscopy data), data defined w.r.t. other modalities (e.g., pictures of artefacts) in different compounds are also (implicitly) linked. Thus, with **multimodal data association** as sketched above, links are not necessarily made explicit but can be virtually defined for certain research tasks when data sets are reused. Given layered compound data arranged in a mesh given by (virtual) links between data (nodes in a mesh), the consortium will deal with the question of how **rankings can be defined using network analysis techniques** in order to provide better support for large NFDI repositories for which answering search queries returns large result sets. In addition to ranking, **features for supporting relevance measures** will be analysed such that a relevant subset of (layered compound) data is identified, i.e., a part of the mesh is identified and offered as the result of, e.g., data retrieval queries. Furthermore, to support interdisciplinary research, **explanation facilities** for relevance and linking itself need to be supported, in combination with respective **data visualization techniques** (which almost always rely on derived data in a compound if materials science data is involved). In our opinion, the above-mentioned requirements define important challenges for which setting up a national research data infrastructure will be fruitful. Insights into solving the above-mentioned data management problems will help researchers as well as the interested public to get access to research data.

Fourth, the consortium will **investigate how collections of (compound) data objects can be managed**. For example, a collection of compound data can be given by a (temporary) collection of physical artefacts (e.g., for an exhibition), by a set of patients involved in a clinical study, by a reference library of PDFs for a particular researcher, or even by a bibliography in a certain paper under investigation for answering a particular research question). While physical collections can be temporary, the data about them will usually persist. A collection is more than the set of objects, and, on the one hand, given that a data object can be part of multiple collections, there is a need to support collection-specific attachments to data objects (e.g., additional layers of interpretation). On the other hand, a collection itself is defined in terms of holistic data involving contributions of all compounds in the collection. In some cases, in the holistic representations particular representations can be identified for single elements of a collection, while in other cases, the holistic representation is indeed only defined in terms of the whole collection (and no part-specific components can be identified). For holistic representations, metadata are required as well.

Last but not least, it should be mentioned that our plan for a NFDI also has the goal to incorporate standards for data structures supporting **analytics about how the previously mentioned data structures are accessed**. Based on analytics data,

scientists from computer science and humanities can optimize algorithms for data linking, data access, and the composition of collections. Researchers from different areas will be supported in **e-science activities** with **provenance** information ensuring **traceability** and **evolution** of entries.

- Brief description of the proposed use of existing infrastructures, tools and services that are essential in order to fulfil the planned consortium's objectives

  We will focus on service provision and competency acquisition to help researchers connect to available interdisciplinary research data of the kind envisioned above, integrated new data, and visualize data in a suitable way for researchers from different areas as well as the interested public. We would like to start in combining research data from the humanities (in particular manuscript research) and natural and materials sciences. This means, images of and texts about artefacts on the one hand, and sensor data, detector data, as well as a huge variety of data from genetics to mineralogy on the other.

  We plan to establish services to keep these data connected in the form of compounds so that in the future a researcher will find relevant data and all relevant associated datasets as well as, for instance, associated publications. This also includes data, publications, and information from other sources on the web. Services should not be limited to national data sources, and researchers should not need to actually know where data are stored.

  The consortium will rely on existing research data infrastructure and will not establish a service for data archiving nor build a new infrastructure. Because of our broad approach we have a very broad variety of data types. It is quite obvious that data types of today may be out-of-date in the future and for sure new data types will arise. Thus, planning for interdisciplinary NFDI structures, one has to be open for new data types.

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

  Because resources are always limited, we have to focus on some areas, and we decided to start with our field of competence in combining humanities and materials sciences research. But we will have an eye on interoperability in general, and so we will develop a kind of framework where it should be possible to pick up standards and tools from other areas. The NFDI-Conference shows common interests with BRIDGE4NFDI, GeRDI4NFDI, ForumX, NFDI Small Disciplines, AI4NFDI, NFDI4HPC, NFDI Web, CompeNDI, NFDI4CS4NFDI and RSE4NFDI and we will build a network with them in the next month.

## 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.
  Data literacy; metadata; API; interoperability; accessibility; openness (closed data are not usable).

- Please indicate which of these cross-cutting topics your consortium could contribute to and how.
  Data literacy, interoperability, accessibility, and openness are cross-cutting topics where we have experience and a strong institutional background. With Universität Hamburg and the Centre for Sustainable Research Data Management (data literacy), the consortium is strongly connected to the “Hamburg Open Science” initiative (openness, accessibility), a joint project of all research institutions in the Free and Hanseatic City of Hamburg.