NFDI4Microbiota - Letter of intent for the National Research Data Infrastructure (NFDI)

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

National Research Data Infrastructure for the Research of Microbiota

Acronym of the planned consortium

NDFI4Microbiota

Applicant institution

ZB MED - Information Centre for Life Sciences
Gleueler Straße 60
50931 Cologne (Köln)
Scientific Director: Prof. Dr. Dietrich Rebholz-Schuman
Managing Director: Gabriele Hermann-Krotz

Spokesperson

Prof. Dr. Konrad Förstner, foerstner@zbmed.de, ZB MED - Information Centre for Life Sciences

3 Objectives, work programme and research environment

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

21 (Biology) and 22 (Medicine)

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

Microbial species (bacteria, archaea, unicellular eukaryotes and viruses) have a strong impact on numerous aspects of human life, starting from health to ecologically relevant processes. Given climate change, our understanding of ecosystems needs to improve dramatically to be able to act against human-made challenges and microbiota have been largely ignored[1]. One of the major medical threats according to the UNO is the rise of antibiotic resistance[2]. Both seemingly uncoupled important issues require a better understanding of the microbial world. Furthermore, countless, not yet cultured species are potential sources of compounds with relevance for biotechnology and medicine, and still need to be explored. One of the biggest challenges in the understanding of microbiota lies in the complexity of the numerous biotic interactions between specific organisms of a microbiome and abiotic environmental factors. Studying individual species and whole

1 Cavicchioli *et al.*, *Nature Review Microbiology*, 2019, https://doi.org/10.1038/s41579-019-0222-5

2 https://www.who.int/who-un/about/amr/en/

microbiomes, in particular mapping and deciphering molecular interactions with their underlying regulatory mechanisms, is a crucial step towards an understanding and possible utilisation of microbial species.

The need for efficient analysis of microbial species and microbiota related data exists in different distinct research fields like biomedicine, agriculture or ocean research. These exemplary research fields have very different scientific questions and origins but share the need for efficient analysis of microbial organisms and the molecular interaction of their members. While high-throughput approaches can easily deliver a plethora of multi-omics data at different molecular levels (DNA, RNA, proteins, metabolites), storing, analyzing and integrating these with available knowledge (e.g. from literature) to decipher the functioning of the individual species and microbiomes as well as their contributions to human health and environmental processes is not trivial. Often researchers do not have the capabilities (computational skills and/or resources) to deal with such data generated by high-throughput technologies and are overwhelmed by the size as well as the complexity of the data and the required processing steps.

The vision of the NFDI4Microbiota consortium is to make the analysis of multi-omics data related to microbial species and diverse microbiomes consistent, reproducible and accessible to all fields of the life science community. It will assist researchers with different scientific challenges to understand individual microbial species and communities as well as the interaction between the species in them. For this purpose, NFDI4Microbiota will provide the computational infrastructure, analytical tools and training for the community to compile, analyze and store various types of data with the aim to decipher microbial species and interspecies interactions on a molecular level. The consortium will enable efficient and reproducible processing of omics data that are generated via high-throughput analysis devices. This includes genomes, transcriptomes, proteomes and metabolomes for individual species as well as the counterpart for microbiota namely metagenomes, meta-transcriptomes, meta-proteomes and meta-metabolomic data. Furthermore, currently emerging analysis approaches for data from single cell sequencing and high-throughput imaging will be supported. Additionally, the consortium will enable to enrich this data by metadata from databases and by knowledge automatically extracted from literature and make the data interoperable. In order to provide a seamless data workflows,

NFDI4Microbiota will collaborate with data generators like sequencing facilities and promote a direct deposition of measured data and connected metadata into the computational infrastructure provided by the consortium. The data will be passed through the analysis pipeline according to the wishes of the researchers who ordered the analysis and raw data, metadata as well as the results will be deposited in repositories for long term availability.

NFDI4Microbiota will fully comply with the FAIR (Findable, Accessible, Interoperable, Re-usable)[3] principles and promote Open Science with all its facets. Sensitive personal data will be treated with necessary care and will undergo anonymization. As part of this, the consortium will define a required, rich set of metadata that describes the sampling conditions and will allow only the submission of data after metadata was provided and quality controlled. Members of our consortium have been leading the development of the International Human Microbiome Standards (IHMS)[4] and NFDI4Microbiota will promote similar standards for other microbiome sources together with the community. Furthermore, the consortium will encourage contributors to choose rather permissive licenses for the submitted data sets in order to avoid legal barriers for data sharing. The consistent management of the submitted data, as well as the contributed rich annotation with metadata form the core to the powerful search of the original data (and given results) and to the efficient reuse and comparison by the research community. In order to generate sustainable solutions, NFDI4Microbiota will follow good software engineering practice to generate a software stack based on FLOSS (Free/Libre/Open Source Software). All developed software will be made public under OSI (Open Source Initiative) compliant licenses.

NFDI4Microbiota will be in continuous exchange with its scientific board and its user council to get feedback and to adjust services to the needs of its research community. Furthermore, established connections with international organisations like ELIXIR and the International Human Microbiome Standards (IHMS) project will be used to align the activities of NFDI4Microbiota with them and find synergies.

3 *https://www.force11.org/group/fairgroup/fairprinciples*
4 *http://www.microbiome-standards.org*

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

- Computational Cloud infrastructure will be provided the German Network for Bioinformatics Infrastructure (de.NBI, spokesperson: Prof. Dr. Alfred Pühler)
- Analysis pipelines will be provided by the European Molecular Biology Laboratory (EMBL, spokesperson: Prof. Dr. Peer Bork), the Center for Biotechnology for the Bielefeld University (CeBiTec, spokesperson: Prof. Dr. Jens Stoye) and the Helmholtz Centre for Infection Research (HZI, spokesperson: Prof. Dr. Alice McHardy)
- Benchmarking frameworks for analysis software and benchmarking data sets will be provided by the Helmholtz Centre for Infection Research (HZI, spokesperson: Prof. Dr. Alice McHardy)
- Reference databases will be provided by the German Collection of Microorganisms and Cell Cultures (DSMZ, spokesperson: Prof. Dr. Jörg Overmann)
- Literature mining based enrichments will be performed by ZB MED - Information Centre for Life Sciences (spokesperson: Prof. Dr. Konrad Förstner)

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

We are in exchange with several other consortia to work on common standards, interfaces and training. This includes but is not limited to the following points:

- With NFDI4Life Umbrella and its sub-consortia we will collaborate on several cross-diciplinary issues of the life science community.
- For standardisation of chemical components such as medication, dietary factor or metabolom data, we will explore the possibilities of standardisation and cross-sectional mapping together with NFDI4Health. Exchange regarding the inclusion of microbiome related data in medical studies is planned.
- We will work with NFDI4BioDiversity on common metadata standards for sampling.
- We will work with NFDI4Agri on common metadata standards for sampling.
- We will cooperate with GHGA on human microbiome data. While GHGA will be responsible for the archiving of the data we will provide tools for the data analysis.

- To facilitate the parallel analysis of microbiota and the immune system, we plan to harmonize metadata and ontologies describing the host as well as identifying and enhancing formalized descriptions of sampling procedures.
- We will work with DataPlant on metadata for omics data and training.
- We will collaborate with NFDI4RSE on standards and recommendations regarding analysis workflows and training.

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.

- Standards regarding high-throughput sequencing data
- Standards regarding data lineage/provenance
- Metadata standards for samples from medical and environmental sources
- Standard procedures for the handling of medical data with privacy issues / need for anonymization
- Solutions for reproducible data analysis workflow
- Infrastructure for ordering, tracking and billing data analyses
- Incentives for data publications
- Training and education

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

We will contribute to the discussion regarding common metadata standards of high-throughput sequencing data and for the sampling from medical and environmental sources.