Research Ideas and Outcomes 9: e110501
doi: 10.3897/rio.9.e110501

Grant Proposal

NFDI4Microbiota – national research data infrastructure for microbiota research

Konrad U. Förstner[‡], Anke Becker[§], Jochen Blom[|], Peer Bork[¶], Thomas Clavel[#], Marius Dieckmann[|], Alexander Goesmann[|], Barbara Götz[‡], Thomas Gübitz[‡], Franziska Hufsky[¤], Sebastian Jünemann[«], Marie-Louise Körner[»], Manja Marz[¤], Ulisses Nunes Da Rocha[ˆ], Jörg Overmann[˅], Alfred Pühler[«], Dietrich Rebholz-Schuhmann[‡], Alexander Sczyrba[«], Jens Stoye[«], Justine Vandendorpe[‡], Thea Van Rossum[¶], Alice McHardy[¦]

‡ ZB MED - Information Centre for Life Sciences, Cologne, Germany
§ Phillips-Universität Marburg, Marburg, Germany
| Justus Liebig University Gießen, Gießen, Germany
¶ EMBL, Heidelberg, Germany
# RWTH University Hospital, Aachen, Germany
¤ Friedrich Schiller University Jena, Jena, Germany
« Bielefeld University, Bielefeld, Germany
» Eppendorf SE, Hamburg, Germany
ˆ Helmholtz Centre for Environmental Research (UFZ), Leipzig, Germany
˅ Leibniz Institute DSMZ German Collection of Microorganisms and Cell Cultures GmbH, Braunschweig, Germany
¦ Helmholtz-Centre for Infection Research (HZI), Braunschweig, Germany

Corresponding author: Konrad U. Förstner (foerstner@zbmed.de)

Reviewable v 1

Received: 03 Aug 2023 | Published: 24 Aug 2023

Citation: Förstner KU, Becker A, Blom J, Bork P, Clavel T, Dieckmann M, Goesmann A, Götz B, Gübitz T, Hufsky F, Jünemann S, Körner M-L, Marz M, Da Rocha UN, Overmann J, Pühler A, Rebholz-Schuhmann D, Sczyrba A, Stoye J, Vandendorpe J, Van Rossum T, McHardy A (2023) NFDI4Microbiota – national research data infrastructure for microbiota research. Research Ideas and Outcomes 9: e110501. https://doi.org/10.3897/rio.9.e110501

Abstract

Microbes – bacteria, archaea, unicellular eukaryotes, and viruses – play an important role in human and environmental health. Growing awareness of this fact has led to a huge increase in microbiological research and applications in a variety of fields. Driven by technological advances that allow high-throughput molecular characterization of microbial species and communities, microbiological research now offers unparalleled opportunities to address current and emerging needs. As well as helping to address global health threats such as antimicrobial resistance and viral pandemics, it also has a key role to play in areas

such as agriculture, waste management, water treatment, ecosystems remediation, and the diagnosis, treatment and prevention of various diseases. Reflecting this broad potential, billions of euros have been invested in microbiota research programs worldwide. Though run independently, many of these projects are closely related. However, Germany currently has no infrastructure to connect such projects or even compare their results. Thus, the potential synergy of data and expertise is being squandered. The goal of the NFDI4Microbiota consortium is to serve and connect this broad and heterogeneous research community by elevating the availability and quality of research results through dedicated training, and by facilitating the generation, management, interpretation, sharing, and reuse of microbial data. In doing so, we will also foster interdisciplinary interactions between researchers. NFDI4Microbiota will achieve this by creating a German microbial research network through training and community-building activities, and by creating a cloud-based system that will make the storage, integration and analysis of microbial data, especially omics data, consistent, reproducible, and accessible across all areas of life sciences. In addition to increasing the quality of microbial research in Germany, our training program will support widespread and proper usage of these services. Through this dual emphasis on education and services, NFDI4Microbiota will ensure that microbial research in Germany is synergistic and efficient, and thus excellent. By creating a central resource for German microbial research, NDFDI4Microbiota will establish a connecting hub for all NFDI consortia that work with microbiological data, including GHGA, NFDI4Biodiversity, NFDI4Agri and several others. NFDI4Microbiota will provide non-microbial specialists from these consortia with direct and easy access to the necessary expertise and infrastructure in microbial research in order to facilitate their daily work and enhance their research. The links forged through NFDI4Microbiota will not only increase the synergy between NFDI consortia, but also elevate the overall quality and relevance of microbial research in Germany.

Keywords

Research data management, FAIR principles, microbiota, NFDI, Germany

List of participants

Centre for Innovation Competence (ZIK) Septomics, de.NBI e.V., Essen University Hospital, Fraunhofer Cluster of Excellence Immune-Mediated Diseases, GEOMAR Helmholtz Centre for Ocean Research Kiel, German Aerospace Center (DLR), Goethe University Frankfurt, Heinrich Heine University Düsseldorf, Helmholtz Centre Potsdam, Helmholtz Institute for RNA-based Infection Research (HIRI), Helmholtz Munich - German Research Center for Environmental Health, Hochschule Anhalt - University of Applied Sciences, Justus Liebig University Giessen (JLU Giessen), Jülich Research Centre, Karlsruhe Institute of Technology, Kiel University, Leibniz Institute for Natural Product Research and Infection Biology - Hans Knöll Institute (HKI), Leibniz Institute of Freshwater Ecology and Inland Fisheries (IGB), Leibniz Institute on Aging - Fritz Lipmann Institute (FLI), Ludwig Maximilian University of Munich, Martin Luther University Halle-Wittenberg,

Max Delbrück Center for Molecular Medicine, Max Planck Institute for Biophysics, Max Planck Institute for Marine Microbiology, Max Planck Institute for Plant Breeding Research, Otto von Guericke University Magdeburg, Robert Koch Institute (RKI), Ruhr-University Bochum, Technical University of Darmstadt, Technical University of Munich, TH Köln - University of Applied Sciences, TIB Leibniz Information Centre for Science and Technology University Library, Ulm University, University Hospital Cologne, University Hospital Erlangen, University Hospital Frankfurt, University of Cologne, University of Greifswald, University of Greifswald, University of Hamburg, University of Jena, University of Kiel, University of Marburg, University of Würzburg (JMU).

1 General Information

Table 1

Table 1.
List of all abbreviations.

API	Application Programming Interface
AU	Administration Unit
BMBF	Bundesministerium für Bildung und Forschung (Federal Ministry of Education and Research)
BoD	Board of Directors
CA	Consortium Agreement
CWL	Common Workflow Language
DataPLANT	Data in PLANT research
de.NBI	German Network for Bioinformatics Infrastructure
DFG	Deutsche Forschungsgemeinschaft German Research Foundation
EBI	European Bioinformatics Institute
ELIXIR	European life-sciences infrastructure for biological information
ELN	Electronic Lab Notebook
ENA	European Nucleotide Archive
EOSC	European Open Science Cloud
FAIR	Findable, Accessible, Interoperable, and Re-usable
GA	General Assembly
GDPR	General Data Protection Regulation
GHGA	German Human Genome-Phenome Archive
INSDC	International Nucleotide Sequence Database Collaboration

MIGS	Minimum Information about a Genome Sequence
MIMAG	Minimum Information about a Metagenome-Assembled Genome
MIMARKS	Minimum Information about a MARKer gene Sequence
MIMS	Minimum Information about a Metagenome Sequence
MISAG	Minimum Information about a Single Amplified Genome
MIxS	Minimum Information about any (x) Sequence
NFDI4Agri	National Research Data Infrastructure for Agricultural Sciences
NFDI4Biodiversity	National Research Data Infrastructure for Biodiversity, Ecology and Environmental Data
NFDI4BIOIMAGE	National Research Data Infrastructure for Biological Imaging and Medical Photonics
NFDI4Cat	National Research Data Infrastructure for Catalysis-Related Sciences
NFDI4Chem	National Research Data Infrastructure for Chemistry
NFDI4DataScience	National Research Data Infrastructure for Data Science and Artificial Intelligence
NFDI4Health	National Research Data Infrastructure for Personal Health Data
NFDI4Immuno	National Research Data Infrastructure for Immunology
NFDI4Ing	National Research Data Infrastructure for Engineering Sciences
NFDI4Life Umbrella	Research Data Management Infrastructure for Life Sciences
NFDI4RSE	National Research Data Infrastructure for Scientific Software
NFDI	Nationale Forschungsdateninfrastruktur National Research Data Infrastructure
NGS-CN	Next Generation Sequencing Competence Network
ORKG	Open Research Knowledge Graph
PUNCH4NFDI	Particles, Universe, NuClei, and Hadrons for the National Research Data Infrastructure
RDA	Research Data Alliance
RDMO	Research Data Management Organizer
RDM	Research Data Management
REST	Representational State Transfer
SAB	Scientific Advisory Board
SRA	Sequence Read Archive
TA	Task Area
TMF	Technologie und Methodenplattform für die vernetzte medizinische Forschung e.V.
UC	Use Case
WMA	World Medical Association

1.1 Name of the consortium in English and German

NFDI4Microbiota - National Research Data Infrastructure for Microbiota Research

NFDI4Microbiota - Nationale Forschungsdateninfrastruktur für Mikrobiota-Forschung

1.2 Summary of the proposal in German

Mikroben - einschließlich Bakterien, Archaeen, einzellige Eukaryonten und Viren - sind von höchster Relevanz für die Gesundheit von Mensch und Umwelt. Das wachsende Bewusstsein für diese Tatsache hat zu einem enormen Anstieg der mikrobiologischen Forschung und Anwendungen in einer Vielzahl von Bereichen geführt. Angetrieben durch technologische Fortschritte in der Hochdurchsatz-Analyse von mikrobiellen Spezies und Gemeinschaften auf molekularer Ebene, eröffnensich der mikrobiologischen Forschung beispiellose Möglichkeiten, um aktuellen und neu entstehenden Herausforderungen zu begegnen. Neben der Bekämpfung globaler Gesundheitsbedrohungen wie der Resistenz gegen antimikrobielle Mittel oder viraler Pandemien, spielt die Mikrobiologie eine Schlüsselrolle in vielen weiteren Bereichen wie Landwirtschaft, Abfallwirtschaft, Wasseraufbereitung, der Sanierung von Ökosystemen, sowie der Diagnose, Behandlung und Prävention von unterschiedlichsten Krankheiten. Diesem breiten Potenzial entsprechend, wurden weltweit Milliarden von Euro in Mikrobiota-Forschungsprogramme investiert. Obwohl sie unabhängig voneinander durchgeführt werden, stehen viele dieser Projekte in engem Zusammenhang. Deutschland verfügt jedoch derzeit über keine Infrastruktur, um die Ergebnisse solcher Projekte zu verknüpfen. Daher gehen potenzielle Synergien verloren bzw. werden nicht realisiert. NFDI4Microbiota wird eine große, heterogene Forschungsgemeinschaft unterstützen, indem es Interaktionen zwischen Forschenden fördert, spezielle Schulungen anbietet und die Generierung, Verwaltung, gemeinsame Nutzung und Wiederverwendung von mikrobiellen Daten und deren Interpretation erleichtert. NFDI4Microbiota wird dies durch eine Vernetzung der Forschungs-Communitys und Training erreichen. Zentral dafür ist die Schaffung eines Cloud-basierten Systems für die Speicherung, Integration und Analyse von mikrobiellen Daten, insbesondere von Omics-Daten, um diese konsistent und reproduzierbar allen Bereichen der Biowissenschaften zugänglich zumachen. Gleichzeitig soll ein reichhaltiges Trainingsprogramm für die breite Masse der mikrobiologischen Forschungsgemeinschaft bereitgestellt werden. Durch die enge Verknüpfung von Ausbildung und Services wird NFDI4Microbiota für eine synergistische, effiziente und somit exzellente Mikrobiologie sorgen.Mit der Bereitstellung einer zentralen Forschungsdateninfrastruktur und Expertise in der Mikrobenforschung wird NFDI4Microbiota zu einem Knotenpunkt für alle NFDI-Konsortien, die mit mikrobiologischen Daten arbeiten, einschließlich GHGA, NFDI4Biodiversity, NFDI4Agri und andere. NFDI4Microbiota wird Nicht-Spezialisten dieser Konsortien einen einfachen Zugang zu mikrobieller Expertise und Infrastruktur verschaffen, um ihnen in Ihren Forschungsbereichen Spitzenforschungzu ermöglichen. Somit werden durch die geschaffene Vernetzung und Infrastruktur nicht nur Synergien zwischen NFDI-Konsortien etabliert und gefördert, sondern wird auch die Gesamtqualität der mikrobiellen Forschung in Deutschland gesteigert.

1.3 Applicant institution

Table 2

Table 2.
Applicant institution and location.

Applicant institution	Location
ZB MED - Information Centre for Life Sciences	Cologne

1.4 Spokesperson

Table 3

Table 3.
Spokesperson and its institutional affiliation.

Spokesperson	Institution, location
Prof. Dr. Konrad Förstner	ZB MED, Gleueler Straße 60, 50931 Cologne

1.5 Co-applicant institutions

Table 4

Table 4.
Co-applicant institutions and their locations.

Co-applicant institutions	Location
Bielefeld Institute for Bioinformatics Infrastructure (BIBI), Bielefeld University	Universitätsstraße 25 33615 Bielefeld
European Molecular Biology Laboratory (EMBL)	Meyerhofstraße 1 69117 Heidelberg
Friedrich Schiller University Jena (FSU Jena)	Fuerstengraben 1 07743 Jena
Helmholtz Centre for Environmental Research (UFZ)	Permoserstraße 15 04318 Leipzig
Helmholtz Centre for Infection Research (HZI)	Inhoffenstraße 7 38124 Braunschweig
Justus-Liebig-University Gießen (JLU Gießen)	Ludwigstraße 23 35390 Gießen
Leibniz Institute DSMZ-German Collection of Microorganisms and Cell Cultures	Inhoffenstraße 7B 38124 Braunschweig
Philipps-Universität Marburg (UMR)	Biegenstraße 10 35037 Marburg
RWTH Aachen University (RWTH)	Templergraben 55 52062 Aachen

1.6 Co-spokespersons

Table 5

Table 5.
Co-spokespersons, their affiliated institutions, and task area(s).

Co-spokespersons	Institution, location	Task area(s) (TA)
Prof. Dr. Alexander Sczyrba & Prof. Dr. Jens Stoye	BIBI, Bielefeld	TA1, 2, 3, 4
Prof. Dr. Peer Bork	EMBL, Heidelberg	TA1, 2, 3, 4
Prof. Dr. Manja Marz	FSU Jena, Jena	TA1, 2, 3
Dr. Ulisses Nunes da Rocha	UFZ, Leipzig	TA1, 2, 3, 4
Prof. Dr. Alice C. McHardy	HZI, Braunschweig	TA1, 2, 3, 4, 5
Prof. Dr. Alexander Goesmann	JLU Gießen, Gießen	TA1, 2, 3, 4, 5
Prof. Dr. Jörg Overmann	DSMZ, Braunschweig	TA1, 2, 3, 4
Prof. Dr. Anke Becker	UMR, Marburg	TA1, 2, 3, 4, 5
Prof. Dr. Thomas Clavel	RWTH Aachen, Aachen	TA1, 2, 3

1.7 Participants

Table 6

Table 6.
Participating institutions and their representatives.

Participating institutions (represented by)	Location
SPP 2141 – CRISPR-Cas functions beyond defense, Ulm University (Prof. Dr. Anita Marchfelder)	Ulm
SPP 2002 – Small Proteins in Prokaryotes, an Unexplored World, University of Kiel (Prof. Dr. Ruth Schmitz-Streit)	Kiel
SPP 2330 - New concepts in prokaryotic virus-host interaction – from single cells to microbial communities, Jülich Research Centre (Prof. Dr. Julia Frunzke)	Jülich
SFB 1371 - Microbiome Signatures, Technical University of Munich (Prof. Dr. Dirk Haller)	Freising
SFB 1021 - RNA viruses: RNA metabolism, host response and pathogenesis, Philipps-Universität Marburg (Prof. Dr. Stephan Becker)	Marburg
EXC 2051 - Balance of the Microverse, Friedrich Schiller University Jena (Prof. Dr. Axel A. Brakhage)	Jena

Participating institutions (represented by)	Location
SFB/TR 124 – Pathogenic fungi and their human host: Networks of interaction, Friedrich Schiller University Jena (Prof. Dr. Axel A. Brakhage)	Jena
Infection Biology and Molecular Biotechnology, Leibniz Institute for Natural Product Research and Infection Biology – Hans Knöll Institute (HKI) (Prof. Dr. Axel A. Brakhage)	Jena
Institute for Artificial Intelligence, Essen University Hospital (Prof. Dr. Folker Meyer)	Essen
Fraunhofer Cluster of Excellence for Immune-Mediated Diseases CIMD (Prof. Dr. Dr. Gerd Geisslinger)	Frankfurt am Main
Chair of General Microbiology, Institute of Microbiology, Friedrich Schiller University of Jena (Prof. Dr. Kai Papenfort)	Jena
Space Microbiology Research Group, German Aerospace Center (DLR) (Dr. Ralf Möller)	Cologne
Department of Biology and Genetics of Procaryotes, Goethe University Frankfurt (Prof. Dr. Jörg Soppa)	Frankfurt am Main
Institute of Medical Microbiology and Hospital Hygiene, Heinrich Heine University Düsseldorf (Prof. Dr. Alexander Dilthey)	Düsseldorf
Cologne Excellence Cluster on Cellular Stress Responses in Aging-Associated Diseases (CECAD), University of Cologne (Prof. Dr. Alexander Dilthey)	Cologne
Institute of Medical Statistics and Computational Biology, University of Cologne (Prof. Dr. Alexander Dilthey)	Cologne
Section 3.7: Geomicrobiology, Helmholtz Centre Potsdam - GFZ German Research Centre for Geosciences (Prof. Dr. Dr. h.c. Reinhard Hüttl)	Potsdam
RNA Biology of Bacterial Infections, Helmholtz Institute for RNA-based Infection Research (Prof. Dr. Jörg Vogel)	Braunschweig
Research Unit for Comparative Microbiome Analysis, Helmholtz Zentrum München German Research Center for Environmental Health (Prof. Dr. Michael Schloter)	Neuherberg
Host Septomics Research Group, Jena University Hospital - Centre for Innovation Competence (ZIK) Septomics (Dr. Tilman Klassert)	Jena
Institute of Bio- and Geosciences – Biotechnology (IBG-1), Jülich Research Centre (Prof. Dr. Michael Bott)	Jülich
Cell function and organ development, Julius Maximilian University of Würzburg (Dr. Markus Engstler)	Würzburg
Department of Bioinformatics, Julius Maximilian University of Würzburg (Prof. Dr. Thomas Dandekar)	Würzburg
Institute for Molecular Infection Biology (IMIB), Julius Maximilian University of Würzburg (Prof. Dr. Cynthia Sharma)	Würzburg

Participating institutions (represented by)	Location
Research Center for Infectious Diseases (ZINF), Julius Maximilian University of Würzburg (Prof. Dr. Cynthia Sharma)	Würzburg
Institute for Virology and Immunobiology, Julius Maximilian University of Würzburg (Prof. Dr. Lars Dölken)	Würzburg
Institute of Microbiology and Molecular Biology, Justus-Liebig-Universität Gießen (Prof. Dr. Gabriele Klug)	Gießen
Institute of Hygiene and Environmental Medicine, Justus-Liebig-University Gießen (Prof. Dr. Linda Falgenhauer)	Gießen
Institute of Medical Microbiology, Justus-Liebig-University Gießen (Prof. Dr. Trinad Chakraborty)	Gießen
Institute of Medical Virology, Justus-Liebig-University Gießen (Prof. Dr. John Ziebuhr)	Gießen
Institute for Applied Biology (IAB), Karlsruhe Institute of Technology (Prof. Dr. Johannes Gescher)	Karlsruhe
Systems Biology and Bioinformatics, Leibniz Institute for Natural Product Research and Infection Biology – Hans Knöll Institute (HKI) (Assoc. Prof. Dr. Gianni Panagiotou)	Jena
Computational and Systems Biology of Aging, Leibniz Institute on Aging – Fritz Lipmann Institute (FLI) (Prof. Dr. Steve Hoffmann)	Jena
Aquatic microbial ecology, Leibniz-Institute of Freshwater Ecology and Inland Fisheries (IGB) (Prof. Dr. Hans-Peter Grossart)	Berlin
Biomedical Center Munich, Ludwig Maximilian University of Munich (Prof. Dr. Nicolai Siegel)	Planegg-Martinsried
Experimental Parasitology, Ludwig Maximilian University of Munich (Prof. Dr. Nicolai Siegel)	Planegg-Martinsried
RG Pharmacognosy, Martin Luther University Halle-Wittenberg (Prof. Dr. Timo Niedermeyer)	Halle
Host-microbiome factors in cardiovascular disease, Max Delbrück Center for Molecular Medicine (Dr. Sofia Forslund)	Berlin
Department of Structural Biology, Max Planck Institute for Biophysics (Prof. Dr. Werner Kühlbrandt)	Frankfurt am Main
Bioprocess Engineering, Max Planck Institute for Dynamics of Complex Technical Systems (Dr. Dirk Benndorf)	Magdeburg
Department of Molecular Ecology, Max Planck Institute for Marine Microbiology (Prof. Dr. Rudolf Amann)	Bremen
Integrative Bioinformatics group, Max Planck Institute for Plant Breeding Research (Dr. Ruben Garrido-Oter)	Cologne
Institute of Microbiology and Genetics, Göttingen University (PD Dr. Heiko Liesegang)	Göttingen

Participating institutions (represented by)	Location
Institute for Medical Microbiology and Hospital Hygiene, Otto von Guericke University Magdeburg (Prof. Dr. Achim Kaasch)	Magdeburg
Robert Koch Institute (Prof. Dr. Lothar H. Wieler)	Berlin
Microbial Biology, Ruhr-University Bochum (Prof. Dr. Franz Narberhaus)	Bochum
Data Science & Digital Libraries Research Group, Technical Information Library (Prof. Dr. Sören Auer & Dr. Irina Sens)	Hannover
Research Group Pharmaceutical Biotechnology, TH Köln - University of Applied Sciences (Prof. Dr. Jörn Stitz)	Leverkusen
Centre for Synthetic Biology, Technical University of Darmstadt (Prof. Dr. Heinz Koeppl)	Darmstadt
ZIEL - Institute for Food & Health, Technical University of Munich (Dr. Ilias Lagkouvardos)	Freising
Infectious Diseases, University Hospital Frankfurt (Prof. Dr. Maria J.G.T. Vehreschild)	Frankfurt am Main
Clinical Microbiome Research Group, University Hospital Cologne (Prof. Dr. Maria J.G.T. Vehreschild)	Cologne
Institute of Microbiology – Clinical Microbiology, Immunology and Hygiene, University Hospital Erlangen (Prof. Dr. Jochen Mattner)	Erlangen
Department of Functional Genomics, University of Greifswald (Prof. Dr. Uwe Völker)	Greifswald
Institute for Biochemistry, University of Greifswald (Prof. Dr. Uwe T. Bornscheuer)	Greifswald
Institute for Microbiology, University of Greifswald (Prof. Dr. Dörte Becher)	Greifswald
Microbiology & Biotechnology, University of Hamburg (Prof. Dr. Wolfgang Streit)	Hamburg
Institute of Clinical Molecular Biology (IKMB), University of Kiel (Dr. Marc Höppner)	Kiel
Life & Medical Sciences Institute (LIMES) - I206, University of Bonn (Prof. Dr. Joachim Schultze)	Bonn

1.8 Names and numbers of the DFG review boards (DFG-Fachkollegien) that reflect the subject orientation of the proposed consortium

[201-01] Biochemistry, [201-05] General Genetics and Functional Genome Biology, [201-07] Bioinformatics and Theoretical Biology, [202-03] Organismic Interactions, Chemical Ecology and Microbiomes of Plant Systems, [203-03] Ecology and Biodiversity of Animals and Ecosystems, Organismic Interactions, [204-00] Microbiology, Virology and Immunology, [205-08] Pharmacy, [205- 31] Clinical Infectiology and Tropical Medicine

2 Scope and Objectives

2.1 Research domains or research methods addressed by the consortium, specific aim(s)

Microbial species such as bacteria, archaea, unicellular eukaryotes and viruses have an immense influence on **every aspect of human life, including health, agriculture, biotechnology, biogeochemical cycles, and climate**. A better understanding of ecosystems is crucial to meeting the broad global challenges posed by human activities, yet the relevance of microbes to addressing these problems has not yet been fully appreciated. According to the UN, one of the greatest threats to human health is the increase in multiple antibiotic resistance in bacterial pathogens. The current COVID-19 pandemic shows the dramatic impact that viral infections can have on every aspect of our lives, while millions of people in developing countries continue to suffer from neglected tropical diseases such as malaria, leishmaniasis, and dengue. At the same time, microbial biomass is a key influencer of atmospheric gas composition and climate. In short, microbes affect every aspect of life on earth, and microbiological research can be directly linked to 8 of the 17 UN Sustainable Development Goals. Countless microbial species are yet to be cultured, many of which could potentially be sources of substances and materials with relevance for biotechnology and medicine. All these issues require a better understanding of the world of microorganisms, and their diversity shows the **impact that any improvements to microbiological research will have on numerous other fields**. One of the greatest challenges to understanding the microbiota lies in the complexity of the numerous biotic interactions between specific organisms within any given microbial community and abiotic environmental factors. The study of individual species and entire communities, in particular the mapping and deciphering of molecular interactions with their functional portfolio and underlying regulatory mechanisms, is a crucial step toward understanding and making more effective use of microbial species.

The activities undertaken by NFDI4Microbiota will support research communities that work with **data on microbial species, including bacteria, archaea, unicellular eukaryotes, and viruses**. Our target group encompasses not only traditional microbiologists, but also **scientists from other fields with microbiological relevance**, such as a clinician who wishes to understand all the factors of a disease, or a biogeochemist exploring how microbes contribute to geochemical cycles.

NFDI4Microbiota will support researchers in handling the following data and associated metadata:

1. omics data including genomics, transcriptomics, proteomics, metabomlics, lipidomics at singlecell, bulk and meta levels, fluxomics, glycomics, epigenomics, phenomics and more,
2. phylogenetic and taxonomic data,
3. physiological and biochemical data,
4. imaging data,

5. systems biological modeling,
6. medical microbiological data,
7. sampling: technical equipment / methods,
8. origin: environment / host / location,
9. utilization (e.g. biotechnological, agricultural applications).

2.2 Objectives and measuring success

NFDI4Microbiota aims to facilitate the digital transformation of the microbiological community in order to significantly accelerate the generation of new, relevant knowledge. Microbiology is intricately linked to numerous other fields, including medicine, agriculture, biotechnology and earth science, so the overall social impact of these improvements will be immense.

To overcome common obstacles to handling data in the field of microbiology (State of the art and needs analysis) NFDI4Microbiota will take steps to achieve the following **key objectives (KOs)**:

- KO-1: Generate a broad **awareness** of the importance of **FAIR principles, open science and reproducible research** in the microbiological community and drive a cultural change toward their widespread adaptation.
- KO-2: Equip the community with the required **skills and literacy** for efficient and data-driven microbial research by providing a **comprehensive training program.**
- KO-3: Increase the **value of other NFDIs** by adding microbial expertise and connecting the national and international microbiology communities.
- KO-4: Improve the research process by **mobilizing, structuring and linking** available data, information and knowledge related to microorganisms.
- KO-5: Support **high-quality research data management** by introducing professional data stewards into the microbiological research process.
- KO-6: Increase the value of data by **standardizing and systematically collecting rich metadata** and building tools for querying.
- KO-7: Make research more reproducible by **standardizing data processing and analysis**.
- KO-8: Provide **computational tools and infrastructure** for the translation of data into new knowledge.
- KO-9: Provide **central information hubs** tailored to address the needs of the microbial research community and its sub-fields.
- KO-10: Continuously **adapt the solutions provided** to meet the future needs of our microbiology community.

Regular user surveys will be conducted to measure our success in meeting these objectives. The various boards will also measure and evaluate a set of **Key Performance Indicators (KPIs)** including completed projects, amount of data collected, data base entries added, access to a central information hub, download and usage of application programming interfaces (APIs), and user help desk requests

3 Consortium

Involvement in other consortia

- Bielefeld University is a co-applicant of NFDI4Biodiversity and PUNCH4NFDI, and a participant of NFDI4Health, FAIRAgro and NFDI4Memory.
- DSMZ is a co-applicant of NFDI4Biodiversity, and a participant of NFDI4Earth.
- EMBL is a co-applicant, and HZI a participant, of GHGA.
- FSU Jena is a co-applicant of NFDI4Biodiversity, and leads NFDI4Chem.
- JLU Gießen is a co-applicant of NFDI4Biodiversity, and a participant of NFDI4Ing and NFDI4Culture.
- UFZ is a co-applicant of NFDIBiodiversity and NFDI4Earth, and a participant of NFDI DeBioData and NFDI4Chem.
- ZB MED leads NFDI4Health, is a co-applicant of FAIRAgro and NFDI4DataScience.

3.1 Composition of the consortium and its embedding in the community of interest

The consortium consists of **ten (co-)applicants** with a broad range of expertise and solid foundations in the national and international microbiology community. Each of the co-applicants conducts their own microbiological research and offers their own infrastructure, though with a different emphasis in each case. The consortium was founded by five members of the NFDI4Life Umbrella consortium. It was subsequently expanded by inviting applications from additional members of the microbiology community. Five of the institutes that applied to become co-applicants were chosen to join the consortium in early 2020. NFDI4Microbiota then invited the community to join as participants. The consortium is supplemented by about 50 participating institutions and networks from all over Germany, including three DFG Priority Programmes, three DFG Collaborative Research Centers and one DFG Cluster of Excellence (see Fig. 1). It also includes the participation of the Next Generation Sequencing Competence Network. Several co-applicants and participants run their own sequencing, mass spectrometry and other high-throughput measurement facilities. The members of the consortium (co-applicants and participants) represent institutions from the different parts of the German science system: universities, the Max Planck Society, the Helmholtz Association of German Research Centres, the Leibniz Association, the Fraunhofer-Gesellschaft and federal government agencies. The members include researchers and infrastructure providers as well as data generators. Furthermore, five national scholarly societies have expressed their support for NFDI4Microbiota and will help shape the activities of the consortium. These are the Deutsche Gesellschaft für Hygiene und Mikrobiologie (DGHM, German Society for Hygiene and Microbiology), the Deutsche Gesellschaft für Parasitologie (DGP, German Society for Parasitology), the Gesellschaft für Virologie e.V. (GfV, Society for Virology), the Vereinigung für Allgemeine und Angewandte Mikrobiologie (VAAM, Association for General and Applied Microbiology) and the Deutsche Gesellschaft für Mykologie (DGfM, German Society for Mycology).

NFDI4Microbiota has already carried out four community workshops and an online survey to get feedback and provide information on the steps it plans to take.

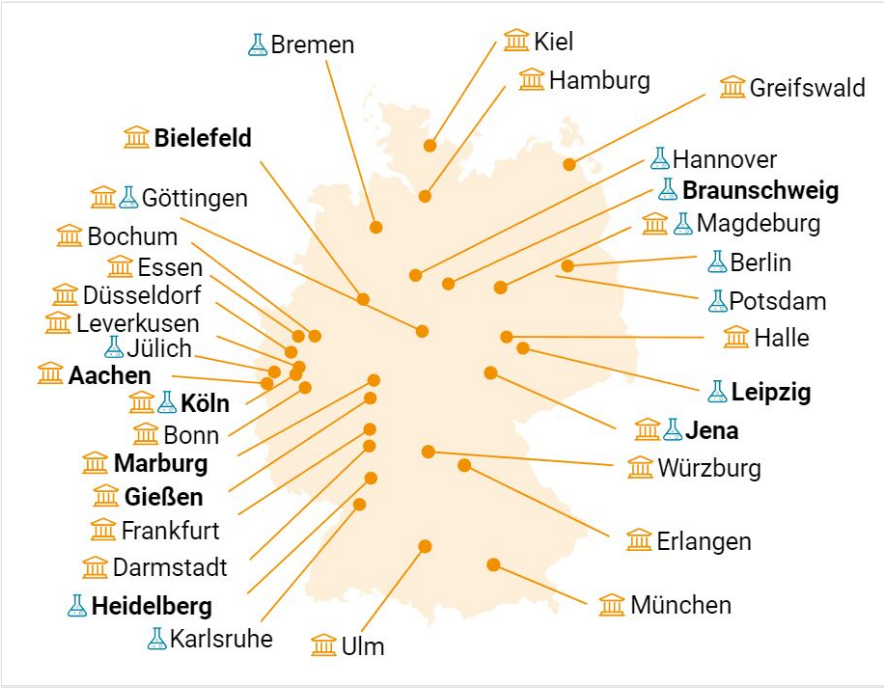

Figure 1. doi

Consortium members in Germany.

The ten NFDI4Microbiota co-applicants are:

Bielefeld Institute for Bioinformatics Infrastructure (BIBI): BIBI is an interdisciplinary academic department of the Faculty of Technology at Bielefeld University. The institute explores the research and service aspects of big data handling in the life sciences. In collaboration with the BMBF-funded “German Network for Bioinformatics Infrastructure” (de.NBI), it helps provide comprehensive, state-of-the-art bioinformatics services and training to users from academia and industry involved in basic and applied life science research. It also seeks to promote cooperation between the German bioinformatics community and international bioinformatics network structures such as ELIXIR. BIBI houses the coordination and administration office of the de.NBI network and the national ELIXIR node ELIXIR-DE. **J. Stoye** (Genome Informatics) is head of the institute and spokesperson for the “Digital Infrastructure for the Life Sciences” graduate school, which forms part of BIBI. **A. Sczyrba** (Computational Metagenomics) heads the Bioinformatics cloud within BIBI and coordinates the de.NBI cloud. He co-organizes the community-driven initiative for the Critical Assessment of Metagenome Interpretation (CAMI), which aims to establish and promote standards in computational metagenomics by creating and organizing benchmarking challenges.

The **European Molecular Biology Laboratory (EMBL)** is Europe's flagship institution for molecular biology, life sciences, and biodata. **P. Bork** is the Director of EMBL Heidelberg (Scientific Activities), which is EMBL's main scientific laboratory, and the site of its headquarters. EMBL is a hub for microbiome infrastructure, both externally, through its mandate within the de.NBI Heidelberg Center for Human Bioinformatics (HD-HuB, directed by P. Bork), its role as data hub in a number of large EU consortia on the human gut microbiome, its major involvement in the GHGA NFDI infrastructure (and specifically in its COVID-19 activities), and its steering committee membership of the International Human Microbiome Consortium, and internally, through infrastructure and training programs supporting microbiome research. Bork is a Senior Group Leader at EMBL, and a leading expert on research of the microbiome and its connection to human phenotypes and diseases, as well as on the global microbiome in the ocean and soil. The Bork lab also provides many bioinformatic web services, averaging approximately one million hits per day. EMBL's International Centre for Advanced Training (EICAT) and its Courses and Conferences program provide dozens of life science training events each year. EMBL also hosts Science and Society programs, the European Learning Laboratory for the Life Sciences (ELLS), and an integrated program ("BioIT") with a mandate for bioinformatics training and community development. Bork collaborates on microbiome infrastructure with its internal partner **EMBL-European Bioinformatics Institute (EMBL-EBI)**, a subsidiary of EMBL, which hosts and operates Europe's largest sequencing data archive (ENA), a microbiome analysis web platform (MGnify), and a central COVID-19 Data Portal, supporting international efforts to advance SARS-CoV-2 research. This direct link, together with the support from EMBL-EBI through Dr. Rob Finn, Team Leader of Microbiome Informatics at EMBL-EBI Hinxton, UK, will secure the sharing of expertise between the EBI and NFDI4Microbiota.

The **Helmholtz Centre for Environmental Research (UFZ, Leipzig)** is one of the world's leading research centers in the field of environmental research. It demonstrates how making sustainable use of our natural resource base – biodiversity, functioning ecosystems, clean water, and intact soils – can benefit both humankind and the environment. Widely regarded as a reliable partner, the UFZ supports policymakers, industry and the general public to better understand the consequences of human actions on the environment and to develop options for social decision-making processes. The UFZ has strong ties with microbiota research and several of their departments and groups will be participating in NFDI4Microbiota. UFZ's Microbial Data Science group (led by **U. Nunes da Rocha**) works on state-of-the-art research data management in the context of research into microbiomes in environmental samples. U. Nunes da Rocha also coordinates the CoLaborative mUlti-domain Exploration of TERRestriAl metagenomes (CLUE-TERRA) consortium that encompasses scientists from 13 institutions and 9 different countries. Further, UFZ will contribute by:

1. generating standards for microbiome analysis in pure cultures (Bioinformatics and Transcriptomics group, **J. Hackermüller**) and complex microbial communities (Microbial Systems Data Science group, U. Nunes da Rocha);

2. integrating multi-omics and meta-organisms (Microbiome Biology group, **N. Jehmlich**; Microbial Systems Ecology group, A. Chatzinotas; Department of Molecular Systems Biology, **M. von Bergen**);
3. integrating imaging of microbiota at nano scales (**N. Musat**); and,
4. bridging human microbiome research to food pathogens, environmental viromes and microeukaryotes (**A. Chatzinotas**). The UFZ has a high-performance computer cluster with over 2,500 nodes that can be used for data handling and analysis of single strains and complex community microbiome data.

Friedrich Schiller University Jena (FSU Jena) has an excellent microbiological scientific network with almost 20 professorships *[1] and an Excellence Graduate School dedicated specifically to microbial communication *[2]. The ”Balance of the Microverse” cluster of excellence, which investigates the dynamic balance of complex microbial communities *[3] has played a significant role in FSU’s research portfolio since 2019. **M. Marz** is full professor for High-Throughput Sequencing Analysis at FSU Jena. She is head of the Bioinformatics Core Facility *[4] at the FSU, founding member, board member and managing director of the European Virus Bioinformatics Center *[5], and founding member of Michael Stifel Zentrum Jena for Data-Driven and Simulation Science *[6]. She is also spokesperson for the H2020-MSCA-ITN-2020 ”Viroinf” *[7] and the ”Digitalization in life sciences” program supported by the Thuringian state government *[8]. **B. König-Ries** works on the distributed, automized usage of resources, i.e. information and functionality, in heterogeneous, dynamic environments, such as microbiomes. One of the key applications for the solutions she develops is research data management, in particular for biodiversity research projects. As well as developing a research data management platform, BEXIS 2, König-Ries is also working on tools to improve data FAIRness and reproducibility of research. **C. Steinbeck** is Professor for Analytical Chemistry, Cheminformatics and Chemometrics. His group develops methods for the computer-assisted elucidation of chemical communication between microbes and other players in ecosystems. He leads NFDI4Chem *[9], the project to build a national research data infrastructure for Chemistry in Germany. **K. Küsel** heads the DFG SFB ”AquaDiva”, which studies community assembly and functioning of groundwater microbiota *[10]. Since 2013, regular joint groundwater sampling campaigns have yielded datasets encompassing taxonomic identification, metabolic activities and functions and environmental parameters. The team combines various profiling technologies including high-resolution DOM analyses, amplicon sequencing, genome-centric approaches, and genome-resolved proteomics and metabolomics. The resulting synergies help to achieve a better understanding of how groundwater functions.

Helmholtz Centre for Infection Research (HZI): The HZI is Germany’s flagship research institute for infection research. Its participation in the Centre for Individualized Infection Medicine places it at the forefront of personalized infection medicine approaches that translate the use of host-pathogen omics data into clinical settings. It plays a leading role in both the German Center for Infection Research (DZIF) and the National Cohort (NAKO). **A. C. McHardy** heads the Department of Computational Biology of Infection Research at HZI. She coordinates the Translational Infrastructure Bioinformatics Services of the DZIF and co-organizes the community-driven initiative for the Critical Assessment of Metagenome

Interpretation (CAMI), which aims to establish and promote standards in computational metagenomics by creating and organizing benchmarking challenges.

Justus-Liebig-Universität Gießen (JLU): The Justus Liebig University Gießen (JLU) was founded in 1607 and is the second-largest university in Hesse. **A. Goesmann** holds the chair for Systems Biology (W3) and is the coordinator of the BiGi Service Center for Microbial Bioinformatics within the BMBF-funded "German Network for Bioinformatics Infrastructure" (de.NBI). He is a member of the DFG-funded Cluster of Excellence Cardio Pulmonary Institute (CPI), the German Center for Infection Research (DZIF), and the Center for Synthetic Microbiology (SYNMIKRO). Currently, A. Goesmann receives funding within the clinical research group on virus-induced lung injury (KFO 309), the transregional collaborative research center on chromatin changes in differentiation and malignancies (SFB TRR81), the research group on communication in host-microbe interaction via exRNA (FOR 5116), the graduate center on regulatory networks in the mRNA life cycle (GRK2355), the LOEWE center for insect biotechnology, the BMBF-funded Computational Life Sciences Initiative, and de.NBI. His group has spent over 20 years working on the development of a modular software platform for the systematic storage, analysis and visualization of very large datasets resulting from high-throughput experiments. The main focus is on DNA sequence analysis, genome annotation (Meyer et al. 2003, Schwengers et al. 2020) and comparative genomics (Blom et al. 2016) including the evaluation of short-read-mapping data (e.g. RNA-Seq, ChIP-Seq), high-throughput analysis of metagenome sequences (Jaenicke et al. 2018), as well as general data management and visualization. Since 2019, **S. Janssen** has held the chair for Algorithmic Bioinformatics (W1, tenure W2). His research focuses on phylogenetic analyses of microbial communities. During his postdoc with Rob Knight (University of California, San Diego), he contributed to the development of open-source software such as the QIIME2 platform and the QIITA study management system. He also specializes in the fields of Algebraic Dynamic Programming (ADP) and RNA secondary structure prediction. A **Bioinformatics Core Facility (BCF)** headed by **B. Linke** is also available at the JLU. This central technology platform provides not only dedicated storage and computing capacities, but also a wide range of bioinformatics software tools and databases and the expertise required to manage large datasets from high-throughput experiments. One of the BCF's key tasks is the structured acquisition and storage of experimental data and the processing of that data with the greatest possible degree of automation. Within de.NBI, the BCF is also responsible for the administration of the cloud computing infrastructure located at JLU. NFDI4Microbiota will also be supported by the Hessian research data infrastructure (HeFDI), which anchors research data management at the participating universities.

Leibniz Institute DSMZ-German Collection of Microorganisms and Cell Cultures: The LeibnizInstitute DSMZ is one of the largest microbiological resource centers in the world. It hosts more than 70,000 biological resources, including 30,000 different bacterial and 5,000 fungal strains. DSMZ operates a large-scale sequencing facility, server structure, and bioinformatics pipelines. It has a track record of bacterial (meta)genome analysis, is a partner of the Genomic Encyclopedia of Bacteria and Archaea (GEBA), and has profound expertise in multivariate analysis and modeling of complex microbial communities. It

recently developed synthetic consortia of strains from the human, mouse and swine microbiome. The DSMZ hosts The Bacterial Diversity Metadatabase (BacDive) – the world's largest database for standardized bacterial phenotypic data – and the List of Prokaryotic names with Standing in Nomenclature (LPSN), a database for tracking changes in the nomenclature of prokaryotes. DSMZ also provides expert knowledge on legal matters, in particular concerning intellectual property and international regulations, including the Convention on Biological Diversity (CBD) and the Nagoya Protocol. **J. Overmann** is the scientific director of the DSMZ and full Professor of Microbiology at TU Braunschweig. He also heads up the DSMZ Department of Microbial Ecology and Diversity research. His research focuses on molecular microbial diversity, bacterial genome evolution, and the molecular basis of bacterial interactions. **B. Bunk** heads the bioinformatics and databases department including the sequencing facility, which primarily focuses on microbial diversity analysis and long read (meta)genome sequencing. His main research interests are large-scale genome assembly, genome dynamics and functional genome analysis. **L. Reimer** is in charge of the database development team and is responsible for the BacDive database. His main research interests are the mobilization and standardization of research data to improve access to and comparability of phenotypic data in microbial research.

Philipps-Universität Marburg (UMR) has a long-standing history and an excellent international reputation in microbiology, which has strongly influenced its structural developments over the past decade. **A. Becker** is professor for Microbial Comparative Genomics and the director of SYNMIKRO, which was jointly established by UMR and the Max Planck Institute for Terrestrial Microbiology (MPIterMIc) in 2010 with a grant from the LOEWE excellence research initiative. SYNMIKRO operates an integrated technology center comprising topical and infrastructural units in Bioinformatics, Structural Microbiology, Mass Spectrometry, Screening and Automation, Fluorescence Microscopy, and Cell Sorting. This center is permanently staffed by experts in state-of-the-art technologies. **M. Lechner** is head of the Bioinformatics unit within SYNMIKRO. He coordinates the Marburg computing cluster (MaRC3a) and the Marburg storage cluster (MaSC). UMR also participates in BioRoboost, an international collaborative network working on synthetic biology standards with a strong focus on microbial systems, and is a co-applicant in NFDI4Biodiversity (headed by Frank Oliver Glöckner) and NFDI4Culture (headed by Torsten Schrade), which offer a number of synergies.

RWTH Aachen University (RWTH): RWTH is one of the leading technical universities in Germany with a focus on engineering and medical sciences. **T. Clavel** is Professor and Head of the Functional Microbiome Research Group at RWTH University Hospital (UKA). The lab combines metagenomic and cultivation approaches to study microbial diversity in the mammalian intestine. It carries out research on the phylogenomy of bacteria and synthetic bacterial communities, omics profiling of the gut microbiome in health and diseases (including technical work and training activities on the improvement of SOPs for sample processing and data analysis), and targeted mechanistic studies in gnotobiotes. T. Clavel participates in several national and international initiatives bringing together experts in microbiology, biotechnology, metagenomics, and informatics, including: (i) the

Euregional Microbiome Centre *[11], a cross-boarder network (Belgium, Netherlands, Germany) enhancing the visibility of microbiome research in the Euregion; (ii) the Microbiota Vault *[12], which aims to conserve the diversity of microbes on earth; (iii) the DFG-funded Collaborative Research Center 1371 *[13]; and (iv) Collaborative Research Center 1382 *[14] located at RWTH. In the latter center, T. Clavel coordinates the core project Integrative Microbiota Analysis that provides standardized wet lab and analysis workflows to the consortium, which is a perfect match for NFDI4Microbiota. T. Clavel's expertise on microbiome research at RWTH is complemented by expertise in bioimaging (**F. Kiessling**), medical informatics and data protection (**R. Röhrig**), and research data management (**L. Bossert**).

ZB MED – Information Centre for Life Sciences (ZB MED): ZB MED provides national services for research data management including DOI services, the PUBLISSO Life Science Repository for publications, and the data management planning tool RDMO4Life, as well as terminology and annotation services. Spokesperson **K. Förstner** heads up the Data Science and Services unit and is responsible for the development of information services, including the LIVIVO discovery service. He also runs a research program with the omics analysis and information research groups, which develop numerous bioinformatical tools, text-mining and linked-open-data solutions. The unit provides data and information literacy training all over Europe and hosts the Regional Coordinator for the DACH region of The Carpentries, a not-for-profit organization that teaches foundational coding and data science skills to researchers and librarians worldwide. **B. Lindstädt** heads the Research Management Group at ZB MED and has tremendous expertise and experience in user consulting in this area. ZB MED is also the applicant institution for NFDI4Health (headed by **J. Fluck**) which offers the potential for numerous synergies.

3.2 The consortium within the NFDI

Microbiology is an important sub-discipline in a number of fields, including medicine, biotechnology, and agriculture. It is therefore important for NFDI4Microbiota to forge close ties with other consortia. The foundations for this are already in place in the form of multiple joint use cases which will be implemented in collaboration with other consortia to put the idea of shared solutions into practice (Fig. 2, M1.7 - Use cases). This approach will ensure NFD to other consortia (see M1.4 - Connection to other NFDI consortia) and work towards interoperability and the sharing of solutions to maximize synergies.I4Microbiota's activities benefit numerous other consortia and, by extension, their research communities. We will explicitly invest resources in forging this connectionto other consortia (see M1.4 - Connection to other NFDI consortia) and work towards interoperability and the sharing of solutions to maximize synergies.

By working with other consortia, we will demonstrate our commitment to building one NFDI rather than several separate ones. Thanks to agreements already made with various consortia to run bilateral and multilateral projects, we will be able to implement this step early on. Several NFDI4Microbiota co-applicants are also involved as co-applicants or participants in other NFDI consortia. This will help ensure an intensive exchange of ideas

and sharing of resources. The cloud infrastructure provided by NFID4Microbiota could serve as a model to be adapted by or federated with other consortia. The consortium's considerable expertise in general data science and data literacy can be shared with others, and our robust training program will cover numerous aspects of general scientific interest.

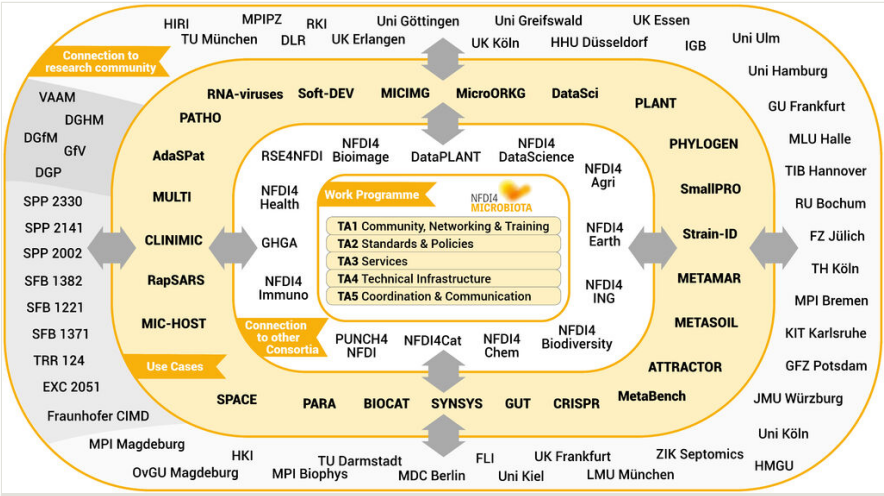

Figure 2. doi

The consortium embedded within other NFDIs linked to the entire community via use cases.

In addition to these specific topics, NFID4Microbiota will work collaboratively with all consortia on cross-cutting topics as a signatory of the Leipzig Berlin declaration on cross-cutting topics in the NFDI (*Leipzig-Berlin-Erklärung zu NFDI-Querschnittsthemen der Infrastrukturentwicklung*) (Bierwirth et al. 2020). For issues of broader relevance – such as a cultural shift in research data management and publication – it is clear that action should be taken by multiple scientific communities working together, rather than individual communities working in isolation. We can contribute expertise to many such topics and are actively involved in communities and initiatives that are driving forward such developments.

3.3 International networking

All co-applicants will help promote the visibility of NFDI4Microbiota and the whole NFDI initiative through their international networks, thereby fostering an exchange of standards and best practice on an international level. We will also implement several concrete solutions together with international partners, working particularly closely with the European Bioinformatics Institute (EBI) and European life-sciences infrastructure for biological information (ELIXIR, Implementation of FAIR principles and data quality assurance). Moreover, we will expand our involvement with the European Open Science Cloud (EOSC), especially EOSC-Life. NFDI4Microbiota has also joined the GO FAIR Microbiome *[15] and GO FAIR Discovery Implementation Network (Open User Interfaces for Increased Visibility of Research Results *[16]) initiatives and will be an active member of the Virus Outbreak Data Network (VODAN) Implementation Network, a joint activity carried out

by CODATA, RDA, WDS, and GO FAIR. Furthermore, the consortium has close ties to the National Microbiome Data Collaborative (NMDC), which is also a member of GO FAIR Microbiome and is a joint venture of the Department of Energy (DOE) National Laboratories, Lawrence Berkeley National Laboratory (LBNL), Los Alamos National Laboratory (LANL), Pacific Northwest National Laboratory (PNNL) and Oak Ridge National Laboratory (ORNL). The NMDC aims to achieve similar solutions for the US as NFID4Microbiota is pursuing for Germany. We also work closely with the J. Craig Venter Institute (JCVI) on the Virus Pathogen Database and Analysis Resource (JCVI). Additionally, we will collaborate with the Beijing Institute of Genomics (BGI) and the Shenzhen Bay Laboratory (SZBL). Based on the positive experiences and ongoing involvement of co-applicants with “The Carpentries”, we will work closely with this not-forprofit organization to develop our training program. The Carpentries offers materials and methodologies (including a train-the-trainer program) for teaching data science skills and data literacy to researchers and people working in library- and information-related roles. Any improvements we make will benefit a global community of trainers and learners.

NFID4Microbiota will also actively extend its international network by running numerous activities (see M1.5 - Connection to international partners).

3.4 Organizational structure and viability

To run operations efficiently, and to give all stakeholders the opportunity to shape the activities of NFDI4Microbiota as a whole, the consortium will implement an organizational structure that consists of several bodies (Fig. 3).

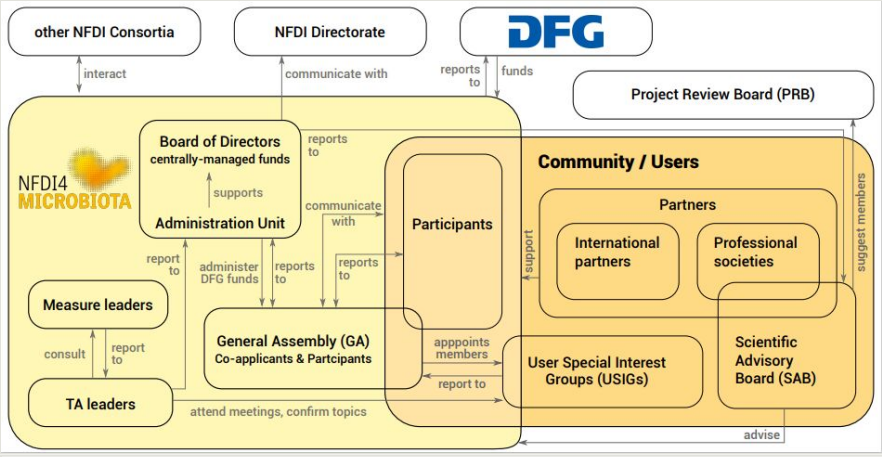

Figure 3. doi

Consortium structure.

Board of Directors (BoD): Although the applicant institution of NFDI4Microbiota is ZB MED, the consortium will be coordinated by a modern, dual leadership team consisting of A. C. McHardy (HZI) and K. U. Förstner (ZB MED), both of whom will represent

NFDI4Microbiota internally and externally. They are responsible for orchestrating the TAs and ensuring the proper connection between participants, use cases and measures. As well as coordinating measures to ensure that milestones are reached and deliverables released on time, they will elaborate solutions with relevant consortium members, should problems arise. They will also be responsible for the dynamic adaptation and growth of the consortium (M1.6 - Sustainability and M5.3 - Dynamic adaptation and growth) and for the achievement of the overall project aims. Finally, they will monitor NFDI4Microbiota's compliance with all NFDI requirements and obligations, and handle legal (M2.4 - Policies and legal issues for data management and reuse), ethical, and contractual issues. The BoD will communicate/meet at least once a month and will report to the General Assembly (GA) and to the Scientific Advisory Board (SAB). They will also have direct contact with the NFDI Directorate.

General Assembly (GA): The GA consists of the ten co-applicant institutions, represented by their respective co-spokespersons named in this application (with voting rights), and of spokespersons of the participating institutions (without voting rights). It is the ultimate decision-making body of the consortium in regard to content and organization. The GA will monitor progress toward meeting the requirements of the TAs and measures. The GA will also ensure NFDI4Microbiota forges connections with other NFDI consortia (M1.4 - Connection to other NFDI consortia), and will communicate with the NFDI committees and the national and international communities (see Task area 5 - Coordination & Communication). ZB MED is responsible for all budget matters and financial decisions made by the GA (see M5.2 - Project financial controlling and reporting) and for any alterations made to the Consortium Agreement (CA) in the future. Finally, the GA is in charge of accepting/rejecting new co-applicants and project proposals by participants, and of deciding on the premature completion or termination of the project. The GA members will communicate/meet at least once a year (the agenda must be made available 3 weeks beforehand) and report to the BoD on progress made within the TAs and measures they lead. Should the consortium – and therefore the GA – grow, a steering committee (SC) consisting of the BoD, task area leaders and/or elected co-applicants will be formed. The SC would take over tasks such as progress monitoring and make strategic decisions.

Administration Unit: This unit will consist of hired staff members that support A. C. McHardy and K. U. Förstner with the coordination of the consortium, including internal communication, milestone and deliverable monitoring, reporting (see M5.2 - Project financial controlling and reporting), quality assurance, and DFG/NFDI administrative tasks. Members of this unit will communicate/meet on a weekly basis and report to the BoD.

Task Area (TA) leaders: TA leaders are in charge of coordinating and managing the work within their TA(s). They are responsible for monitoring of the overall workflow within their TA(s) and scientific supervision of the measures contained in their TA(s). They achieve this by regularly consulting with the measure leaders (i.e. co-applicants responsible for the implementation of the work assigned to their measures), who are themselves responsible for reaching milestones on time, solving problems occurring at the operational level, and reporting to the TA leaders. TA leaders will communicate/meet once every six months and report to the BoD on TA progress and on potential risks to the achievement of project aims.

User Special Interest Groups (USIGs): The User Special Interest Groups represent users' interests in specific issues and will help to bring users' voices into the decision-making process. USIGs consist of experts on a specific subject who represent the research community for defined cross-cutting topics in technical, scientific or social domains. Members of the USIG are appointed through the submission of nominations or self-nominations, which are decided on by the GA. USIG meetings are coordinated and supported by a member of the consortium, who will usually be a TA leader. This TA leader will typically attend USIG meetings, together with a member of the coordination team. Each USIG will elect a spokesperson. Where necessary, USIGs can suggest inviting experts from outside the consortium as temporary or permanent guests. New USIG topics can be proposed by any representative of the community and initiated based on the decision of the GA, who can also terminate USIGs. The following USIGs will be implemented in the initial phase:

1. Training & Education
2. Tools & Services
3. Data Quality/Maturity & FAIR Data Principles
4. Infrastructure & Storage
5. Increasing Diversity
6. Ambassador Program Development

Information flow between the consortium and USIGs is bi-directional; that is, USIG members serve to disseminate information into the communities while also conveying information to the consortium. Each USIG will communicate/meet at least once a year and report its recommendations to the GA.

Scientific Advisory Board (SAB): The SAB comprises scientific experts and representatives of the professional societies (and as such representatives of the user community) from the fields of microbiology and bioinformatics. The SAB members will be involved in research data infrastructures and international initiatives with direct relevance to NFDI4Microbiota. We will favor potential SAB members who work in related international infrastructures or consortia in order to facilitate the exchange of expertise among these entities. The SAB will provide strategic input and advise the consortium on 1) the research data management strategy, 2) proposals and plans, and 3) the needs of the research community.

This body will also highlight critical issues and emerging global trends. The SAB will thus support the further development of NFDI4Microbiota (see M1.6 - Sustainability and M5.3 - Dynamic adaptation and growth). SAB members will communicate/meet at least annually and suggest strategic planning ideas to the BoD.

Project Review Board (PRB): The PRB will be responsible for selecting new projects within the scope of the flexible funding and allocation mechanism (M5.3 - Dynamic adaptation and growth), thus reacting to new and unforeseen scenarios. It will be staffed based on suggestions made by the SAB and decision-making will include a peer review process.

Management of funds: the BoD will be responsible for the financial management of the grant (M5.2 - Project financial controlling and reporting). They will be in charge of establishing the legal framework with all NFDI4Microbiota co-applicants by setting up a Consortium Agreement (CA). They will also administer the financial transactions of the grant from the DFG to the consortium members in accordance with the cooperation agreement, the grant agreement (Bewilligung) and the usage policy (Verwendungsrichtlinie) of the DFG. Finally, they will monitor the use of funds distributed to the consortium members according to the work program set out in this application. Each co-applicant will be given an assigned budget and will be responsible for the proper allocation of this budget to fulfill the duties and obligations defined in this application. The spokesperson / coordinator will be responsible for organizing the disbursement of centrally managed/internal funds (i.e. flex funds). All co-spokespersons and participants are eligible for these funds. Should a co-applicant leave the consortium, the GA will also be responsible for reallocating funds.

All data hubs and several contributing infrastructures (e.g. the Quantitative Biology Center, all de.NBI Cloud nodes, GNC data infrastructure) have corresponding usage regulations. These regulations need to be harmonized across partners to ensure transparent and uniform processes.

3.5 Operating model

NFDI4Microbiota will operate on the basis of 1) the Consortium Agreement (CA) which will be laid down before the funding period, 2) the organizational structure of (co-)applicants and participants (see Organizational structure and viability), and 3) a centrally established legal framework (”NFDI e.V.”) which is currently under construction.

NFDI4Microbiota will be able to build on the experience ZB MED gained from developing NFDI4Health, in terms of complying with the German Fiscal Code and Value Added Tax Act, as well as in a broader sense. Moreover, NFDI4Microbiota will work closely with the NFDI Directorate to tailor its Consortium Agreement to general recommendations.

NFDI4Microbiota aims to include further funded and non-funded members and has allocated funds and an organizational process to do this. Furthermore, NFDI4Microbiota will strive to extend its reach internationally, particularly in conjunction with EOSC-Life.

4 Research Data Management Strategy

4.1 State of the art and needs analysis

Our selection of key objectives (Objectives and measuring success) and the measures to achieve them is founded on the collective experience of all ten co-applicants based on their diverse roles in the microbiology research community. It is also derived from the feedback gathered during our four NFDI4Microbiota community workshops (held in Nov. 2019, Mar. 2020, Jun. 2020, Aug. 2020), as well as from an online survey we conducted.

The picture that emerged is of a microbiology research community – both in Germany and beyond – that would benefit hugely from a broad and consistent application of FAIR principles and open science approaches. This community also has a clear need for more and better training. Over 90% of those surveyed think that greater integration of public data into their research would be beneficial.

The use of new high-throughput measurement devices has driven a huge increase in the amount of data generated in all fields of microbiology (e.g. Fig. 4a - Growth of metagenomic sequencing data in SRA over time). Yet the use and reuse of this data, information, and knowledge has failed to keep pace with these developments and fallen far short of its potential. Many groups working on experiments in the field are already struggling to handle and store omics datasets reliably, and metadata tends to be of poor quality due to the often patchy or belated efforts made to collect it. It is even possible that Germany has a comparatively weak knowledge-sharing culture: the usage statistics of the Sequence Read Archive (SRA) reveal that Germany submits a smaller amount of sequencing data per researcher than other European countries with similar research funding and expertise (Fig. 4b).

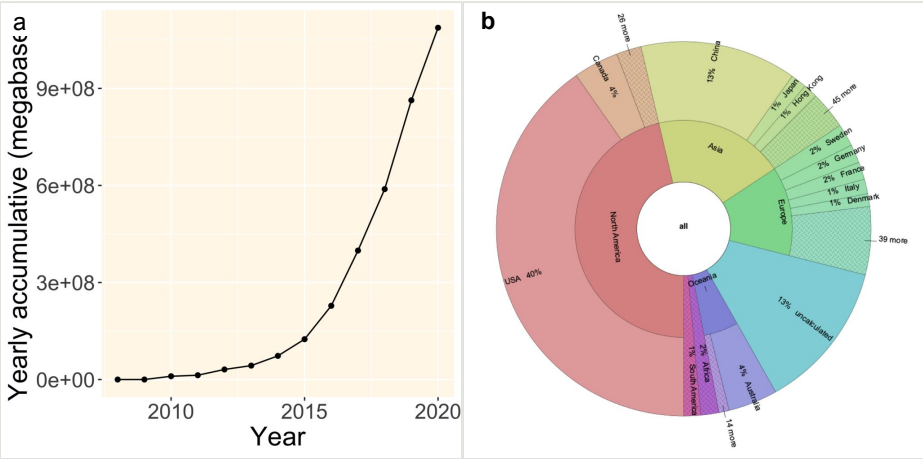

Figure 4.

Growth and distribution of metagenomic datasets in SRA.

a: Growth of metagenomic sequencing data in SRA overtime. doi
b: Contribution to SRA *[17] by location. doi

In addition to improvements in the handling of omics datasets, the community would also like to see solutions for smaller datasets, most of which are manually curated. These data often stem from projects with a limited funding period and cannot subsequently be reused.

But it is not just the storage and handling of data and metadata per se that many members of the community consider to be challenging and time-consuming. They also struggle to analyze omics data due to the need for external expertise and a frequent lack of transparency and reproducibility. Though solutions that provide formal workflow

descriptions do exist (e.g. Common Workflow Language), they are by no means widely employed by the community and are not required by most academic journals.

The lack of options for sharing data processing workflows and results with a larger community leads to a serious waste not only of precious human time, but also of computational resources.

Microbiology has also failed to experience widespread adoption of semantic Linked Open Data approaches. Though still very much under development, this powerful set of technologies has attained a certain level of maturity in recent years, enabling data, information, and knowledge to be linked and queried across multiple research domains. By offering new methods of interactive searching, these technologies pave the way for discovering new insights and lie at the heart of numerous leading tech companies, including Google *[18].

Furthermore, there is a significant lack of infrastructure and procedures for restricting access to data. Sensitive data plays a role in several sub-fields of microbiological research, as well as in related fields and medical applications, for example. Such data may include microbiome studies from human samples or projects investigating host and pathogen interactions of clinical isolates.

In summary, on the one hand we have a broad set of available solutions that have the potential to boost microbial research immensely but, on the other, a situation where these solutions are not part of common practice in the community. Based on our experience and feedback from the community, it appears that this failure to exploit the potential of existing solutions is largely due to the following **obstacles**:

1. A lack of awareness of the benefits of FAIR principles in the community,
2. Deficiencies in digital literacy among researchers,
3. Poorly structured data, information, and knowledge,
4. A lack of appreciation for good research data management and the allocation of insufficient human resources,
5. Poor standardization and usage of metadata,
6. A lack of consent for data processing and analysis,
7. The absence of central information points for the communities,
8. Rapid development of new technologies requiring constant adaption on different levels, and
9. The time required to communicate on both national and international levels.

All our key objectives (Objectives and measuring success) try to help the community to overcome these obstacles.

4.2 Metadata standards

NFDI4Microbiota will manage access to a defined set of primary data, including microbiology-related omics data and data from related methods, e.g. from structural

biology. Such data are subject to existing standards and are submitted in well-defined formats such as FASTQ and BAM. We will further include derived data resulting from computational processing, such as assemblies, metagenomeassembled genomes (MAGs) and functional profiles of metagenomics data. These will be stored in established, machine-readable formats to facilitate further processing and statistical analysis.

Metadata and their formats, on the other hand, are not at all well-defined. Metadata itself is an imprecise term used differently depending on the application. We here define metadata as anything that offers further information on data stored via the NFDI4Microbiota platform, such as provenance, generation, processing or context (e.g. clinical or environmental data).

NFDI4Microbiota will increase compliance with standards and clarify connections between the numerous loosely coupled metadata formats that are relevant for the microbiological community. These cover the different processing stages, from collection of the biological samples to the final analysis result. Among others, we will include standards and checklists for sample descriptions, such as the MIxS standards, developed by the Genomic Standards Consortium (Field et al. 2008). Such standards also demonstrate how different data-type-dependent formats can be united in one framework, with MIxS including MIGS for genomes, MIMS for metagenomes, MIMARKS for marker genes, and MISAG for single amplified genomes. Recently, MIxS checklists have been extended to analytically derived entities, such as MIMAG for genomes assembled from metagenomes. Furthermore, these standards have been extended to cover the context from which a sample came, with 17 official “environmental packages” developed, such as for the built environment (Yilmaz et al. 2011), human skin, soil, and hydrocarbon resources. Metadata will be divided into technical data (according to the technologies used for data generation), biological data (depending on the analyzed type of microbiota), and environmental data (according to environmental variables). The definition of standards for each metadata type will be established by specialists from that research area within the consortium, e.g. data type specialists for technical metadata (M2.1 - Data & metadata standards).

With regard to the formalization of data analysis workflows, we will utilize popular workflow execution environments such as the Common Workflow Language (Amstutz et al. 2016), Snakemake, Nextflow and Cromwell. To maximize synergies, the consortium will closely collaborate with international partners and platforms to explore their suitability for application in the NFDI4Microbiota platform and make use of standardized analysis workflows wherever possible. Along with the standardized acquisition of the corresponding metadata, this approach will facilitate reproducibility and reusability of results generated by NFDI4Microbiota workflows.

4.3 Implementation of FAIR principles and data quality assurance

To ensure successful implementation of the FAIR principles and open science concepts in the microbiology research community, we will increase awareness and offer training and education. In addition, we will provide infrastructure, analytical services, and individual support for putting these approaches into practice. By helping researchers to achieve high-

quality data and metadata management during the entire research cycle and generating tangible additional value for the community and individual researchers, we will continuously drive the required cultural change towards FAIR and open science. To this end, we will collaborate closely with the GO-FAIR initiative as mentioned in M1.5 - Connection to international partners.

To provide a seamless data flow from start to end, we will actively work on all stages including the experimental phase, data generation, and analysis, as well as the deposition and publishing of data together with analytical workflows and results (Fig. 5). The four Next Generation Sequencing Competence Network Centers established by the DFG are participants in NFDI4Microbiota, and several other consortium participants and co-applicants have their own facilities for high-throughput data generation (sequencing, mass spectrometry and imaging). We will work together with them to establish workflows for capturing metadata. Furthermore, we will implement our technical solutions in close collaboration with the European Bioinformatics Institute (EBI), a subsidiary of EMBL, which is a partner institution in this consortium. Data will be submitted directly to their dedicated databases (e.g. ENA *[19] for sequencing data, PRIDE *[20] for proteomics data, Metabolights (Haug et al. 2020) for metabolomics data, BioImage Archive (currently under development) *[21] for images) and to other ELIXIR Deposition Databases for Biomolecular Data *[22], but not made public until requested by the submitting researcher. NFDI4Microbiota will accept all data without constraints. However, the platform will offer workflows for quality control as well as storing corresponding results together with the data, thus enabling users to select high-quality datasets independently (M3.6 - Long-term preservation). Additionally, we will compile/generate and store high-quality exemplary datasets, taking into account community feedback and needs, to enable users to familiarize themselves with data and data types for common purposes. As EBI is a member of the International Nucleotide Sequence Database Collaboration (INSDC) *[23], data will also be mirrored to generate redundancy. Metadata deposited at EBI can be further improved in a curation process that will be supported by us and the research community. The metadata will mostly be stored as entries in the BioSample and BioProject registries, two standards developed by NCBI and also used by ENA. This will also allow us to further link and extend the data entries to other research items.

Data analysis is an essential step in the research process, yet researchers often fail to provide a formal description of the steps performed during analysis. In terms of the cloud-based analytical services we will be providing, we will make use of the Common Workflow Language (Amstutz et al. 2016) as an open executable formal description, and use the CWLProv (Khan et al. 2019) format to represent workflows, including their output artifact in order to formalize data provenance. As part of the EOSC-Life Workflow Hub *[24], a registry of scientific workflows is currently under development which will be used to make the workflows compliant with FLOSS and FAIR principles. Furthermore, the pipeline will also be made available in a containerized manner.

We will work long-term on generating bundles comprising all the information from a project, building on the Research Object project *[25], which aims to encapsulate all relevant information in a machinereadable fashion.

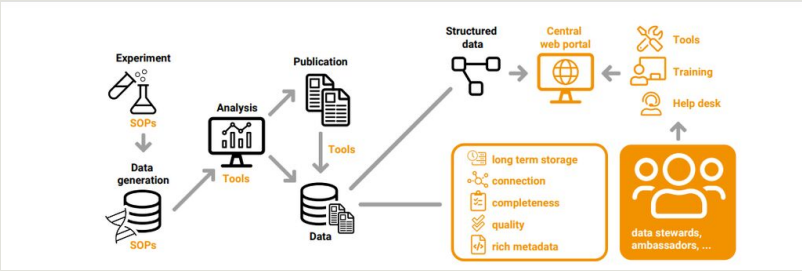

Figure 5. doi

Flow of data from generation to deposition and reuse.

4.4 Services provided by the consortium

As its centerpiece, the consortium will host the **NFDI4Microbiota Hub** (M3.1 - Central web portal), offering the community a central platform through which they can access data and services. The consortium already has a web presence *[26] that provides information on microbiome-related topics and promotes events such as the previously hosted community workshops and user questionnaires. This will be further developed into the **NFDI4Microbiota Training Program** (M1.1 - Training and education), a joint educational program for the community that will include a range of events such as workshops, summer schools, and hackathons as well as providing online teaching material. Targeting different groups ranging from students and doctoral researchers to postdocs, it will cover general topics such as the importance of research data management, but also specialist topics such as specific microbiota analyses and tool usage. The consortium members have solid expertise in training, and can thus build on and extend existing materials and infrastructure to reach a broader audience. Further, the consortium is planning a sophisticated **NFDI4Microbiota Communication Program**, which will involve outreach to the different NFDI4Microbiota communities (M1.3 - Community outreach and public relations), other NFDI consortia (M1.4 - Connection to other NFDI consortia) and international partners (M1.5 - Connection to international partners). It will aim to leverage the various partners' existing connections and create new ones. In particular, NFDI4Microbiota has established the **NFDI4Microbiota Ambassador Program** (M1.3 - Community outreach and public relations) to guarantee close links and easy communication with its participants.

The local ambassadors in the participating groups, which have largely already been appointed, will receive dedicated training on NFDI4Microbiota matters and will act as mediators to facilitate communication between participants and the consortium. On a more functional side, the consortium will provide the **NFDI4Microbiota Storage System** and the **NFDI4Microbiota Service Programme**, both integrated into the NFDI4Microbiota Hub (see Fig. 6). For the storage system, the consortium can build on existing infrastructure provided by consortium members as well as synergies to e.g. the EMBL-EBI infrastructure. Similarly, the consortium members are already developing and benchmarking tools and software for microbiome-related research, which can be built on and integrated into NFDI4Micobiota. By working together to unify and extend these existing efforts within the

NFDI4Microbiota infrastructure, we will reach out and support a diverse community of microbiome scientists and create immense additional value on the path towards synergistic and reproducible national and international microbiome research.

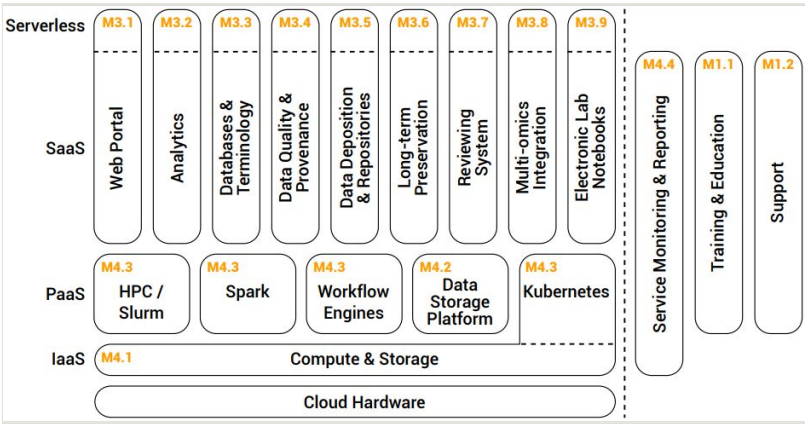

Figure 6. doi

Composition of the technical infrastructure and supporting activities.

5 Work Programme

Table 7

Table 7.

Overview of the task areas, their measures, and responsible co-spokesperson(s).

Task area	Measures	Responsible Co - Spokeperson(s)
1. Community, Networking & Training	**M1.1** - Training and education **M1.2** - Support **M1.3** - Community outreach and public relations **M1.4** - Connection to other NFDI consortia **M1.5** - Connection to international partners **M1.6** - Sustainability **M1.7** - Use cases	Alice McHardy (HZI), Konrad Förstner (ZB MED)
2. Standards & Policies	**M2.1** - Data & metadata standards **M2.2** - Experimental procedure standards **M2.3** - Workflow standards **M2.4** - Policies and legal issues for data management and reuse	Peer Bork (EMBL), Jörg Overmann (DSMZ)
3. Services	**M3.1** - Central web portal **M3.2** - Analytical services **M3.3** - Databases & terminology services **M3.4** - Data quality & provenance services **M3.5** - Data deposition & repositories **M3.6** - Digital preservation **M3.7** - Reviewing / commenting system for data **M3.8** - Systems biology, modeling, and multi-omics integration **M3.9** - Electronic lab notebook (ELN)	Alexander Goesmann (JLU), Manja Marz (FSU Jena)

Task area	Measures	Responsible Co - Spokeperson(s)
4. Technical Infrastructure	**M4.1** - Computational infrastructure operations **M4.2** - Data storage platform **M4.3** - Development of software tools & common components **M4.4** - Service monitoring & reporting	Alexander Sczyrba (BIBI), Alexander Goesmann (JLU)
5. Coordination & Communication	**M5.1** - Project governance **M5.2** - Project financial controlling and reporting **M5.3** - Dynamic adaptation and growth	Konrad Förstner (ZB MED), Alice McHardy (HZI)

5.1 Task area 1 - Community, Networking & Training

Coordinators: HZI (lead), ZB MED (co-lead)

NFDI4Microbiota will establish a cross-cutting infrastructure, supporting its target communities with access to microbiome-related data, analysis services, data and metadata standards, and training.

As such, the success of NFDI4Microbiota hinges on actively involving the target communities, both in defining key aspects of the established infrastructure, as well as in its active use. This task area is therefore dedicated to engaging the diverse NFDI4Microbiota target communities and building and strengthening a network of mutual support and constructive feedback. To this end, we will establish an active system for training and educating (M1.1 - Training and education), which will involve workshops and training events for the communities, as well as documentation of past events, in particular the archiving of training events and conferences, including available material. NFDI4Microbiota will further establish a system to support participating institutions in the process of submitting and analyzing data using the platform (M1.2 - Support). We will reach out to the diverse NFDI4Microbiota target audiences via various channels (e.g. active participation in important conferences, mailing lists, social media presence, interaction with DFG-funded research consortia), thus promoting the NFDI4Microbiota mission and usage, and fostering community acceptance (M1.3 - Community outreach and public relations). We will further explore and actively promote synergies to existing and future NFDI consortia, develop joint use cases and organize joint events (e.g. conferences and workshops), in part by leveraging the connections of applicant institutions who are already part of other NFDI consortia (M1.4 - Connection to other NFDI consortia). Similarly, we will connect to international partners to further develop microbiome standards and ensure they are also compatible on an international level (M1.5 - Connection to international partners).

Finally, as this is a considerable challenge for all infrastructures, we will dedicate time and effort to making NFDI4Microbiota a sustainable platform, and guaranteeing its long-term usage and maintenance (M1.6 - Sustainability).

Fig. 7

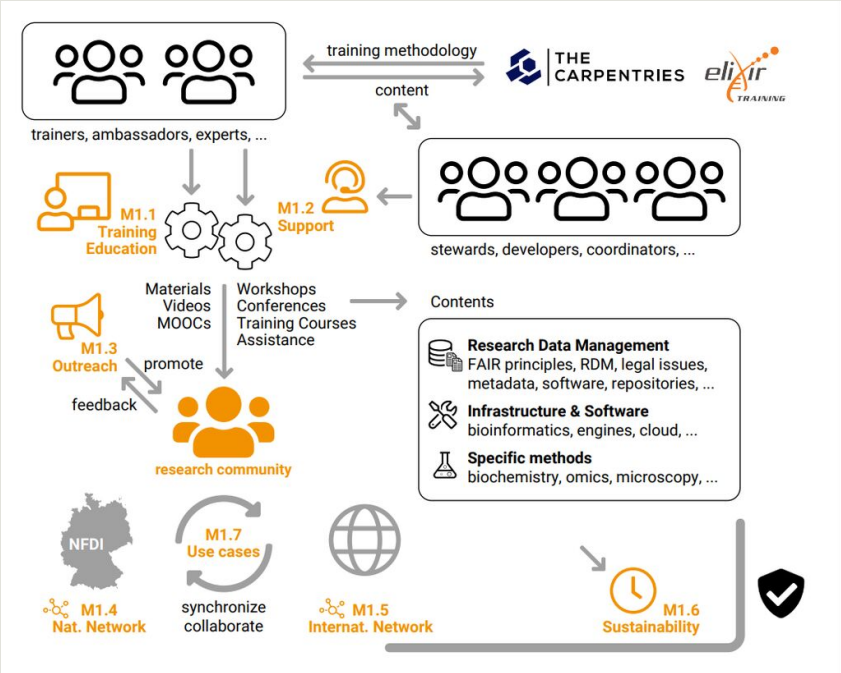

Figure 7. doi

Overview of measures in Task Area 1.

NFDI4Microbiota has already been joined or supported by many different scientific societies, networks and international partners, which represent and provide links to a very large and diverse target community. NFDI4Microbiota collaborates with:

- **Other NFDI consortia** (DataPLANT, NFDI4Agri, NFDI4Biodiversity, NFDI4BIOIMAGE, NFDI4Cat, NFDI4Chem, NFDI4DataScience, NFDI4Earth, GHGA, NFDI4Health, NFDI4Immuno, NFDI4Ing, NFDI4Life Umbrella, NFDI4RSE, and PUNCH4NFDI): collaborations to leverage synergies and reach common target audiences.
- **Priority Programme (SPP) 2141** (A. Marchfelder): research network focusing on ”CRISPR-Cas functions beyond defence”.
- **SPP 2002** (R. Schmitz-Streit): research network focusing on ”Small Proteins in Prokaryotes, an Unexplored World”.
- **SPP 2330** (J. Frunzke): research network focusing on ”New Concepts in Prokaryotic Virus-host Interactions – From Single Cells to Microbial Communities”.
- **Collaborative Research Centre (SFB 1021)** (S. Becker): research institution focusing on ”RNA viruses: RNA metabolism, host response and pathogenesis”.
- **SFB/Transregio (TR) 124 - FungiNet** (A.A. Brakhage): research institution focusing on ”Pathogenic fungi and their human host: Networks of interaction”.
- **SFB 1371** (D. Haller): research institution focusing on ”Microbiome Signatures - Functional Relevance in the Digestive Tract”.

- **SFB 1382** (T. Clavel): research institution focusing on ”Gut-liver axis - Functional circuits and therapeutic targets”.
- **Cluster of Excellence (EXC) 2051** (A.A. Brakhage): research network focusing on ”Balance of the Microverse”.
- **Fraunhofer Cluster of Excellence CIMD** (G. Geisslinger): research cluster focusing on ”ImmuneMediated Diseases”.
- **German Society for Hygiene and Microbiology (DGHM)** (G. Häcker): supporting society representing 1,900 medical scientists, microbiologists, hygiene specialists and clinicians.
- **German Society for Parasitology (DGP)** (M. Engstler): supporting society representing 457 parasitologists active in zoology, botany, medicine, microbiology, hygiene, veterinary medicine, plant protection or pest control.
- **ELIXIR Germany** (A. Tauch): ELIXIR Node in Germany and Europe that provides bioinformatics infrastructure for life science research and works on service, training and cloud computing.
- **German Mycological Society e.V. (DGfM)** (M. Thines): supporting society representing 1,400 mycologists.
- **Society for Virology (GfV)** (R. Bartenschlager): supporting society representing about 1.000 virologists.
- **Association for General and Applied Microbiology (VAAM)** (C. Lang): supporting society representing 3,500 microbiologists.
- **Wikimedia Deutschland e.V.** (F. Heine): non-profit organization dedicated to establishing and promoting the creation, collection, and distribution of free knowledge in all parts of society.
- **BGI-Shenzhen** (J. Li): international partner focusing on microbiome and pathogenome.
- **BioRoboost** (M. Porcar Miralles): international partner that aims to foster synthetic biology standardization through international collaboration.
- **Brazilian Microbiome Project (bmp)** (V. Pylro): international partner that aims to assemble a Brazilian Metagenomic Consortium/Database
- **Euregional Microbiome Center (EMC)** (J. Penders): international partner dedicated to crossing borders between disciplines in microbiome research in the Euregional landscape.
- **European Molecular Biology Laboratory - EBI** (EMBL-EBI) (R.D. Finn): international partner that runs MGnify, the largest metagenomics resource of publicly analyzed microbiome data.
- **FAIR Microbiome Implementation Network** (S.L. Martin): international partner that works with microbiome research communities to accelerate the discovery process.
- **Genome Institute of Singapore** (N. Nagarajan): international partner dedicated to developing new approaches for metagenomic analysis and modeling microbial communities.
- **GO FAIR Discovery IN** (P. Kraker): international partner that provides interfaces and other userfacing services for data discovery while exploring innovative ways of enabling discovery.

- **GO FAIR VODAN** (B. Meerman): international partner whose purpose is to make virus outbreak data FAIR, both for humans and machines.
- **Harvard T. H. Chan School of Public Health** (C. Huttenhower): international partner dedicated to understanding microbial community function and developing new computational methods for human and microbial systems biology.
- **French National Institute for Food, Agricultural and Environmental Sciences (INRAE)** (J. Doré): international partner focusing on the generation and analysis of meta-omics data. • J. Craig Venter Institute (R.H. Scheuermann): international partner engaged in developing the Virus Pathogen Resource (ViPR).
- **DOE Joint Genome Institute (JGI)** (N.C. Kyrpides): international partner focusing on microbiome data integration and analysis.
- **Million Microbiomes from Human Project (MMHP)** (Y. Liu): international partner that aims to construct a microbiome map of the human body and build the world's largest open-access database of the human microbiome.
- **National Centre of Competence in Research (NCCR)** Microbiomes (J.R. van der Meer & J. Vorholt): international partner that aims to understand the unifying principles of microbiome functioning, to develop tools to diagnose microbiome status, and to devise strategies to intervene and restore imbalanced microbiomes.
- **National Institutes of Health (NIH)** (G. Storz): international partner studying bacterial transcriptomes and identifying/characterizing small, regulatory RNAs and proteins.
- **Open Science MOOC** (C.C. Erdmann): community of over 1,000 researchers and practitioners interested in and committed to helping equip students and researchers with the skills they need to excel in a modern research environment.
- **Shenzhen Bay Laboratory (SZBL)** (L. Li): international partner dedicated to understanding the fundamental biology underlying health and diseases.
- **The Carpentries** (K.L. Jordan): international partner focusing on teaching foundational coding and data science skills to researchers worldwide.
- **The Microbiota Vault, Inc.** (M.G. Dominguez-Bello): global non-profit initiative dedicated to conserving the tremendous microbial diversity within microbiota.

This TA will leverage and extend existing ties, ensuring that the whole of the microbiome community is reached, trained and included.

5.1.1 M1.1 - Training and education

Contributors: BIBI (lead, 62 PM), ZB MED (co-lead, 36 PM), UFZ (28 PM), FSU Jena (24 PM), JLU (11 PM), UMR (36 PM), EMBL (40 PM), RWTH (44 PM), HZI (30 PM), DSMZ (30 PM)

Goals: To coordinate, organize and document training events in NFDI4Microbiota to create a comprehensive training program. To promote online courses and online material, and to raise awareness of the value of the FAIR principles and appropriate strategies for RDM.

Training, i.e. the organized process of educating a target audience such that each individual is thereafter able to perform a particular task on their own, is a key means by

which NFDI4Microbiota will connect with the scientific community, which in turn is the main objective of all NFDI efforts. Providing an extensive and broad scope of high-quality training is consequently one of the main measures of NFDI4Microbiota, as highlighted by the extensive contributions of each co-applicant. One of the first tasks will be the establishment of a single platform on which all training events provided by the network will be announced. This platform will be integrated into the central NFDI4Microbiota Hub (M3.1 - Central web portal) and, where possible, also link to comparable platforms of other NFDI initiatives. This platform will also hold documentation of past events, in particular by providing an archive of workshops, conferences, and training courses, including any materials available online. As training is a multiplier for building community trust in NFDI and for extending outreach, events will be actively promoted in collaboration with the WGCO (M1.3 - Community outreach and public relations) via e.g. the central Hub, social media and mailing lists. Our goal is to gradually reach out to specialist research centers (e.g. “Microbiome Signatures” SFB1371, “Gut-liver axis” SFB1382, or “Microverse” EXC2051), societies (e.g. the German Society for Hygiene and Microbiology (DGHM), the Association for General and Applied Microbiology (VAAM), the German Society for Virology (GfV)), and other consortia and universities. Once the announcement platform for past and future events is in place, we will work on an online training platform that enables organizers and experts to upload slideshows and online training materials, and to prepare full online courses on dedicated topics. This will allow researchers with different levels of knowledge to carry out their own off-schedule training and prepare themselves for courses with higher prerequisites.

The type of training events may vary, depending on the topic. They can be performed in person or virtually, in the form of workshops or summer schools. Some may include practical sessions such as hands-on lessons or hackathons. However, we will encourage organizers and presenters to provide online teaching materials as Open Educational Resources (OER) on hosting platforms like GitHub or ReadTheDocs and, once established, we will fully integrate these into the NFDI4Microbiota online training platform and archive them for the long-term. Additionally, and where practical, videos of the actual presentations will be recorded and added to the collection of online materials.

The training topics will cover different scopes and levels of detail, thereby addressing different user profiles, from students to group leaders, as well as different communities. The scope will range from general issues of research data management, data quality, bioinformatics and cloud accessibility to more NFDI4Microbiota-specific courses including e.g. microbial analyses, metagenomics and proteomics. The provision and management of FAIR data and the leverage of the FAIR principles, as well as the important aspect of experiment documentation and metadata standards, will be persistent and recurring themes throughout all training events.

In table 8 an overview of a potential first-year portfolio of NFDI4Microbiota training topics, covering specific training activities of which several are existing, upcoming events or are planned for the near future is provided. This overview also highlights the readiness of each co-applicant to substantially contribute to the overall training and education efforts with further resources. Consequently, each coapplicant will, on average, organize or

significantly contribute to four specific training events per year, irrespective of whether these are in-person, virtual, or new online courses. In order to mitigate topic redundancy and temporal overlaps, the coordinator will actively support partners in creating and offering joint training events, consolidating the partners' expertise to achieve the best training experience and knowledge transfer. Additionally, a large number of courses will be given in cooperation with other NFDI networks (e.g. NFDI4Biodiversity, NFDI4DataScience, NFDI4Health or GHGA) and institutes or research networks (e.g. DZIF), and coordinated closely with other national, European and international relevant training activities, e.g. through the de.NBI training platform and ELIXIR TeSS. Table 8

Table 8.

Current and planned portfolio of NFDI4Microbiota training topics.

Scope	Topics	Organizer/ Contributors/ Experts(TA)
Infrastructure & Software	• General cloud introduction (e.g. de.NBI) • General bioinformatics training • Workflow engines (e.g. Galaxy, Snakemake, Nextflow, CWL)	• BIBI, JLU • EMBL (BioIT), JLU, HZI, FSU Jena, ZB MED • JLU
Research Data Management	• Foundational programming and data science skills (e.g. Software, Data and Library Carpentries workshops) • Metadata standards & FAIR principles • Cloud compute platform services (e.g. Kubernetes) • Public repositories - depositing and searching	• EMBL, ZB MED (DACH coordinator) • ZB MED, UFZ, FSU Jena, DSMZ • BIBI, JLU • UFZ, ZB MED
Domain-Specific Training	• Metagenomics (16S & WGS) • Transcriptomics & metatranscriptomics • Microbial and viral bioinformatics, pathogens & evolution • Genomics & annotation • Proteomics, MS, crystallography, microscopy, chemoinformatics & biochemistry • Multi-omics integration of data	• RWTH, TUM, UFZ, EMBL, BIBI, HZI, JLU, FSU Jena • EMBL, UMR, BIBI, HZI, FSU Jena • JLU, HZI, FSU Jena • UMR, JLU, RWTH • UMR, UFZ • EMBL, BIBI, UFZ

NFDI4Microbiota will also coordinate a mentorship program, where mentors will advise mentees on scientific and career matters. Mentees will be paired with mentors of the appropriate field and career stage from different institutes. Mentors and mentees will be recruited directly from participant and coapplicant institutes, as well as through attendance at training courses and via a sign-up system on the online web portal (M3.1 - Central web portal). Virtual training courses will guide mentors and mentees on how they can each contribute and gain the most from the program. A suggested schedule of virtual meetings will be provided, as well as suggested topics for discussion. The mentorship program will

be open to all, but extra effort will be made to include and support individuals from underrepresented groups in science. This program will enhance connectivity throughout the NFDI4Microbiota community and support one-on-one training.

To guarantee high-quality training and education, we will strive to implement three additional measures. First, we will develop general good practice guidelines for training, most probably in the form of a white paper. This will help training organizers to set up events and help contributors to use tried-and-tested solutions for their presentations and FAIR data sharing. Second, together with M1.3 - Community outreach and public relations and M4.4 - Service monitoring & reporting, we will establish a training feedback form to gather feedback in a consistent and comparable manner (analogous to the de.NBI training surveys). This will allow us to respond directly to feedback, report back to organizers, and check the overall quality of the training based on user assessments. Third, we will work closely with The Carpentries, an international not-for-profit organization that teaches foundational coding and data science skills to researchers and librarians worldwide *[27]. It offers a detailed, evidence-based teaching methodology, including programs to train instructors and maintain OER teaching materials. Members of NFDI4Microbiota are heavily involved in this community and the consortium will build upon these connections. Moreover, we will collaborate with ELIXIR throughout the “Train the Trainer program” (TtT) and plan to establish at least one annual event to communicate the good practice guidelines more directly. This also has the potential to serve as a certificate for NFDI4Microbiota trainers, providing credits toward scientists’ careers. Furthermore, we will collaborate with Open Science MOOC *[28] to work on common content and formats. We will also invite external teaching experts to evaluate the training program and expand our content and methodologies based on their feedback. Table 9

Table 9.

Milestones to be achieved in measure 1.1 - Training and education.

Milestone	Month	Description
MS1.1.1	3	Training events calendar launched
MS1.1.2	6	Training feedback system launched
MS1.1.3	12	Mentorship program launched
MS1.1.4	12	Initial guidelines released on good training and mentorship
MS1.1.5	12	First annual training cycle established, organized by co-applicant
MS1.1.6	14	First annual “Train the Trainer” event cycle established
MS1.1.7	18	Training feedback system finalized
MS1.1.8	24	Initial release of online training platform
MS1.1.9	24	Guidelines on good training and mentorship finalized
MS1.1.10	48	Final release of online training platform

5.1.2 M1.2 - Support

Contributors: ZB MED (lead, 30 PM), DSMZ (co-lead, 30 PM), UMR (12 PM), all other co-applicants (sig. in-kind)

Goals: To offer a broad spectrum of support to overcome individual problems faced by researchers including the writing of data management plans, data submission, metadata collection, and legal issues (e.g. data licenses). To help with the FAIRification of databases and provision of dedicated software tools.

Operations: The support team will consist of a help desk maintained by data stewards who can assist directly with the majority of requests. The help desk will use a ticket system to forward requests to the partner that is most qualified to offer support in the respective area, e.g. the specialists in training and education (M1.1 - Training and education) or in services (Task area 3 - Services). NFDI4Microbiota will also reach out to other NFDI consortia to enable requests to be forwarded to them if they have the required expertise and, vice versa, to receive requests from their user community.

Online community: A help forum will be established as part of the NFDI4Microbiota Hub (M3.1 - Central web portal) that will enable users to support and interact with each other, solve common issues on their own, and enliven the community with discussions on the most suitable strategies for solving common challenges. Users can post issues of specific or general interest, which can be answered either by other users or NFDI4Microbiota moderators. Over time, this will lead to the development of an extensive knowledge resource. NFDI4Microbiota ambassadors (M1.1 - Training and education) will carry this multiplier effect into their local institutions and foster use of the forum. Private issues can be directed to our central, multi-level help desk and will be tracked via an internal ticketing system (MS1.2.1). To enable help desk agents to thoroughly answer user questions, we will compile comprehensive, welldocumented catalogs of available analytical services – both those hosted by consortium members and others – together with collections of SOPs (M2.2 - Experimental procedure standards) and best practices (as e.g. benchmarked in M3.2 - Analytical services). Agents might point users to suitable upcoming training events (through our central calendar), online tutorials and matching forum threads or suggest use of electronic lab notebooks hosted by NFDI4Microbiota.

Data and metadata management: In order to ensure the high quality of data and metadata, NFDI4Microbiota will support participating institutions and other members of the community at all stages of their research. This also includes the structuring of metadata and referrals to further experts, if needed. Especially in the projects that will be performed together with the participants (M1.7 - Use cases), the data stewards will be tightly embedded in data management, helping to generate data management plans, providing advice on metadata collection and helping with the actual data and code deposition.

Dataset/database/software FAIRification: The support team will also assist researchers during all the phases of a database's life cycle, focusing on **FAIRificatio**n, safeguarding of data, research software and knowledge (M3.3 - Databases & terminology services).

Established databases can be further improved e.g. by search machine optimization, standardization of data, extensive linking, and web service development. Integration of databases into the central web portal (M3.1 - Central web portal) will be encouraged and actively supported. Old datasets that are hard to reuse due to technical or legal issues can also be rescued and made FAIR, either by generating simple machine-readable files or by including the data on Wikidata *[29] or on the Open Research Knowledge Graph (ORKG) *[30]. ZB MED and TIB have started a Microbiology Observatory for this purpose in ORKG.

Evaluation: The support activities will also be regularly evaluated to develop new services (Task area 1 - Community, Networking & Training), standards, training materials (M1.1 - Training and education), or other activitiesThe support activities will also be regularly evaluated to develop new services (Task area 1 - Community, Networking & Training), standards, training materials (M1.1 - Training and education), or other activities. Table 10

Table 10.
Milestones to be achieved in measure 1.2 - Support.

Milestone	Month	Description
MS1.2.1	9	Help desk established
MS1.2.2	12	Online forum launched
MS1.2.3	18	Procedure established to translate request evaluations into new training and services
MS1.2.4	24	Procedure developed to exchange requests with other NFID consortia
MS1.2.5	40	Compilation of information resources established

5.1.3 M1.3 - Community outreach and public relations

Contributors: HZI (lead, 30 PM), ZB MED (co-lead, 24 PM), FSU Jena (7 PM), all other co-applicants (sig. in-kind)

Goals: To engage a diverse target community, provide information on NFDI4Microbiota's mission and highlight the advantages of FAIR and standardized data sharing for data holders and platform users. Active community engagement will ensure that platform functionalities and services are in line with user expectations and needs, thus promoting active use and community contributions. To this end, NFDI4Microbiota will establish a Working Group for Community Outreach (WGCO), responsible for organizing and overseeing community outreach and engagement and promoting consortium activities.

The WGCO will include the lead and co-lead of this measure, as well as the lead for Service Monitoring & Reporting (M4.4 - Service monitoring & reporting). Its tasks will be i) to inform and promote NFDI4Microbiota activities via various channels, ii) to coordinate community engagement via workshops, surveys, and training, and iii) to gather and summarize feedback from platform users, such that it can be used as part of efforts to improve the platform by the technical task areas responsible for data storage and service

implementation (M4.1 - Computational infrastructure operations, M4.2 - Data storage platform, M4.3 - Infrastructure software components).

Informing, promoting and gathering feedback: NFDI4Microbiota already operates a general website *[26], offering information on services to engage the community and providing state-of-the-art information on microbiome-related data and their analysis. This site will be expanded into a central portal that will act as the point of entry to all other NFDI4Microbiota resources and services, such as data upload and download, cloud-based analyses and workflows, and training courses (M3.1 - Central web portal). The site will additionally feature informative videos and regular podcasts on microbiomerelated topics. Furthermore, NFDI4Microbiota will promote its activities to different user communities on social media (e.g. Twitter), through its established mailing list, in press releases, and via other channels. Four community workshops have already been hosted to communicate and discuss NFDI4Microbiota's mission. The feedback from these events and further questionnaires has been incorporated into this proposal. NFDI4Microbiota will continue to implement community feedback from various channels, such as the USIGs, the SAB and further questionnaires.

NFDI4Microbiota Ambassador Program: NFDI4Microbiota will establish the NFDI4Microbiota Ambassador Program to create a direct connection to participant institutions. The aim is to engage young researchers (PhD students or postdoctoral researchers) as local contacts at participating institutes and other institutes interested in the program. Local ambassadors will get dedicated training in the core topics relevant to research data management and the services provided by NFDI4Microbiota. They can communicate what they learn in local internal meetings, colloquia and individual meetings. They also represent an additional point of contact with NFDI4Microbiota users. For the institutions, ambassadors can serve as experts to consult on NFDI4Microbiota topics. All participating institutions have agreed to provide at least one person to act as an ambassador.

Data providers: One of the main NFDI4Microbiota goals is to promote synergistic and responsible data sharing and analysis, and to clarify the value of these processes for data holders and analysts. To foster acceptance among data generators, it is important to guarantee and also communicate that data owners will keep control over their data and get to decide when their data will be released to the public. We will also closely interact with key data contributors such as the Next Generation Sequencing Competence Network (NGS-CN) to ensure that NFDI4Microbiota accepts and complies with their wishes and needs. Together with the different user communities, we will identify the most relevant datasets that should be represented in the platform, as well as required analysis features (M3.2 - Analytical services).

Engaging the communities: To promote and strengthen the consortium and its goals, several events with different objectives are planned. An annual international meeting will gather NFDI4Microbiota partners with the diverse user communities, presenting an opportunity for direct interaction between consortium members and other scientists. The

SAB will be part of this annual meeting to strengthen its ties to the community, so that it is aware of and can thus advise on possible concerns and wishes.

Microbiology-related communities: NFDI4Microbiota's target audience includes the diverse set of microbiology-related and application-related communities of e.g. bacteriologists, virologists, clinician scientists, epidemiologists, parasitologists, mycologists, and microbial ecologists who focus on clinical, environmental, agricultural and biotechnological applications. Among these users are the key data contributors and expert users who need to be informed of NFDI4Microbiota use cases and encouraged to deposit data and use the analytical services, as well as give feedback. To reach the individual groups, the consortium will be present at important community conferences, such as the Annual Meeting of the German Society for Hygiene and Microbiology (DGHM) and Association for General and Applied Microbiology (VAAM), the Annual Meeting of the Society for Virology (GfV), the International Virus Bioinformatics Meeting (EVBC), the Scientific Conference of the Mycological Society (DGfM), the DGEpi-Online-Jahrestagung for epidemiologists and the Annual Meeting of the German Society for Parasitology (DGP), which will be extended to include e.g. events focusing on biotechnology, agriculture, clinical research, environment and ecology. Here, posters will be presented or an information booth provided to promote NFDI4Microbiota topics and enable direct contact with consortium representatives, and discussion of needs. Links to existing consortia, community activities, such as participating SFBs, SPPs and EXCs will be leveraged. The different societies will also be addressed individually at the annual NFDI4Microbiota event, with contributions and workshops specific to their application scenarios. Further communication of the NFDI4Microbiota mission and goals will be achieved through societyspecific newsletters and mailing lists, as well as social media posts. This measure will be connected to M4.4 - Service monitoring & reporting, ensuring that user wishes are communicated and rapidly implemented into the platform where applicable.

Biobanking community: Jörg Overmann from DSMZ will leverage his contacts with the biobanking community and promote NFDI4Microbiota's goals on FAIR data sharing, as well as active use of the platform. We will explore synergies with the different biobanks, such as linking biobanked material and metadata with corresponding data deposited via the NFDI4Microbiota infrastructure.

Data science, bioinformatics and computational biology communities: This target audience, similar to the diverse audience of microbiologists, will be informed of NFDI4Microbiota use cases and encouraged to deposit data and use the analytical services, as well as to provide feedback as representatives of the technical side of microbiome research. The bioinformatics target audience will also be addressed to identify or further add to the core services offered on the NFDI4Microbiota platform.

Engaging citizen scientists: To include a more general audience, the NFDI4Microbiota Hub (M3.1 - Central web portal) page will include contributions designed for a more general public, featuring informative videos on the consortium and user surveys. Further outreach will be realized via press releases, newsletters, blog posts, videos and podcasts, and events such as a microbiology science slam.

Harmonizing multiple communities: To establish a successful and harmonized NFDI4Microbiota infrastructure used by various target communities, their requirements need to be unified as far as feasible and sensible. To achieve this, we have defined, and will further extend, use cases (M1.7 - Use cases) of interest to multiple communities. NFDI4Microbiota will closely collaborate and work in line with existing European and international organizations, such as ELIXIR, EOSC-life and other NFDI consortia (M1.4 - Connection to other NFDI consortia) to harmonize data access on a global and long-term basis. Table 11

Table 11.
Milestones to be achieved in measure 1.3 - Community outreach and public relations.

Milestone	Month	Description
MS1.3.1	4	NFDI4Microbiota conference/workshop presentations established (e.g. for VAAM/DGHM); six or more to be held per year
MS1.3.2	6	Bimonthly press releases introduced on activities and aims of NFD4Microbiota
MS1.3.3	9	NFDI4Microbiota Ambassador Programme launched
MS1.3.4	12	First annual user group meeting, including sessions aimed at non-specialist scientists; two or more further events to be held per year
MS1.3.5	13	Annual Community workshops, Microbiology Science Slam, and semiannual Ambassador training established
MS1.3.6	13	Publication of annual report on community feedback and its realization established

5.1.4 M1.4 - Connection to other NFDI consortia

Contributors: HZI (lead, 27 PM), ZB MED (co-lead, 6 PM), FSU Jena (6 PM), UMR (3 PM), JLU (2 PM), BIBI (4 PM), EMBL (3 PM), UFZ (3 PM), DSMZ (sig. in-kind)

Goals: To make NFDI an effective and sustainable endeavor, the different consortia need to closely collaborate and work in a trans-disciplinary manner. NFDI4Microbiota will closely collaborate with others, align activities, and interconnect services (The consortium within the NFDI).

Cross-cutting topics: As a signatory to the Leipzig/Berlin declaration on NFDI cross-cutting topics (“Leipzig-Berlin-Erklärung zu NFDI-Querschnittsthemen der Infrastrukturentwicklung”) NFDI4Microbiota will actively join the discussion and implement shared topics. We will contribute to common solutions by sending representatives to inter-consortia conferences, workshops and working groups, and assist in organizing such events. The outcome of these discussions will be included in our activities.

Bilateral/multilateral collaboration with other consortia: We will explore synergies and harmonize activities with other consortia on all levels, e.g. training activities, achieving compliance, and further extending existing data and workflow standards. This will be

facilitated by the fact that several NFDI4Microbiota co-applicants are also active within other NFDI consortia and joint activities are already planned. Implementing joint use cases (M1.7 - Use cases) serves to generate common solutions and achieve interoperability between solutions from different consortia. We will continuously work on new use cases that help develop common standards as well as their technical implementations. We will also work with GHGA, DataPLANT, NFDI4Biodiversity, and the international community to further develop training portfolios as well as metadata and workflow standards.

Synergies with NAKO, GHGA and NFDI4Health: To further leverage existing ties between NFDI and the epidemiology community, we will work together with NFDI4Health and NFDI4Medicine. NFDI4Microbiota will work with the NAKO (Nationale Kohorte) health study, which aims to investigate the causes of chronic diseases through extensive sampling of the German population and by generating different types of omics data, including microbiota-related data types. GHGA is already collaborating with NAKO in the analysis of human omics data and NFDI4Microbiota will explore synergies with both consortia.

To this end, we will generate and promote informative material on FAIR data sharing and the advantages this offers, thereby responding to and alleviating the fears and obstacles currently hindering rapid and FAIR data deposition practices. One of the main aims here is to ensure that future data deposition will occur in data repositories that follow FAIR principles and standards established by NFDI4Microbiota and other NFDI consortia. Table 12

Table 12.
Milestones to be achieved in measure 1.4 - Connection to other NFDI consortia.

Milestone	Month	Description
MS1.4.1	4	First strategic meeting between NFDI4Microbiota and other NFDI consortia representatives
MS1.4.4	6	Organization of and participation in joint conferences and workshops to facilitate exchange between established consortia
MS1.4.2	10	Selection of flex fund projects realizing shared use cases jointly implemented with other NFDI consortia (e.g. Use case 'MIC-HOST': Integration of host genetics and microbiome data)
MS1.4.3	14	Concept finalized for shared data deposition and data exchange workflows with interacting NFDI consortia, e.g. GHGA, NFDI4Agri and NFDI4Health

5.1.5 M1.5 - Connection to international partners

Contributors: FSU Jena (lead, 30 PM), UFZ (co-lead, 17 PM), UMR (3 PM), JLU (2 PM), RWTH (in-kind)

Goals: To connect NFDI4Microbiota with the international scientific landscape, maximize synergies and determine how to best integrate efforts across all areas, e.g. standards, analytical workflows, data and result sharing, as well as training efforts.

We will collect feedback to define key concepts in the various measures, such as M2.1 - Data & metadata standards, M3.7 - Reviewing/commenting system for data, M3.3 - Databases & terminology services, M2.2 - Experimental procedure standards. In addition, training courses offered by NFDI4Microbiota will be devised in collaboration with international experts (M1.1 - Training and education).

Working groups: Leading national and international scientists will be invited to engage with the NFDI4Microbiota network to discuss emerging topics, foster creativity, and identify cutting-edge trends in microbiota research. They will also receive support for their own analyses from NFDI4Microbiota. The working groups will meet six times in the first year, three times in the second year, and then once a year after that. This timeframe can be adjusted to reflect emerging topics during the funding period.

Example 1: To develop a database for bacteriophages, we need to develop a new data structure. We make agreements with NCBI, EMBL/EBI, and ViPR database developers to discuss this in detail in a three-day workshop in month 6.

Example 2: Human microbiome research advances at an unprecedented pace. To keep up with the cutting-edge science in the field, we create a working group linked to the latest developments in standard operating procedures designed to optimize data quality and analytical workflows for processing human microbiomes, thereby increasing comparability. HZI, EMBL and UFZ (all co-applicants of NFDIMicrobiota) would be involved in such a working group, together with the ELIXIR metagenomics community and international NFDI4Microbiota partners, such as Curtis Huttenhower (Harvard), Rob Finn (EMBL-EBI), Nikos Kyrpides (JGI), Niranjan Nagarajan (A Star Genomics Institute Singapore), the BGI, and Rob Knight (UCSD).

Dedicated events: We will organize a five-day kick-off meeting (one week, partially remotely) including international partners as keynote speakers. It will be attended by all NFDI4Microbiota co-applicants and participants and other target-community scientists who express interest. This event will be combined with the first meetings of the different working groups. Subsequently, we will organize international conferences in Germany annually, with international partners and other key experts. Table 13

Table 13.

Milestones to be achieved in measure 1.5 - Connection to international partners.

Milestone	Month	Description
MS1.5.1	2	Kick-off meeting
MS1.5.2	6	Working groups established
MS1.5.3	12	Annual international conferences established

5.1.6 M1.6 - Sustainability

Contributors: HZI (lead, 12 PM), EMBL (co-lead, 10 PM), FSU Jena (3 PM)

Goals: To develop and implement strategies for establishing NFDI4Microbiota as a long-term hub for microbial research in Germany.

All infrastructure and service projects face the challenge of how to continue operations after the original funding phase. To enable NFDI4Microbiota services to be offered to the community over the long term, NFDI4Microbiota will study different ways of applying them. The modular and service-oriented nature of the planned NFDI4Microbiota infrastructure will facilitate long-term maintenance, and thus sustainability. We will investigate the acquisition of funding from the federal government and/or states for the entire infrastructure and for individual components, such as the data storage infrastructure or analysis toolbox. We will also explore common solutions with other NFDI consortia. To ensure visibility, we will promote NFDI4Microbiota's mission and progress to decision-makers and funding bodies, such as the Federal Ministry of Education and Research (BMBF) and the German Council for Scientific Information Infrastructures (RfII) *[31], which focuses on the topics "Research Data - Sustainability - Internationality" and represents a broad range of scientific disciplines, actors and institutions. Among other topics, this council provides recommendations on development of the NFDI, publishing discussion papers and giving recommendations on establishing successful NFDI consortia *[32]. The council has the role of a consultant on policy and science, but can also initiate discussion processes. To update stakeholders and decision-makers on NFDI4Microbiota proceedings, we will provide annual reports including challenges faced and progress made. In the planned community activities (M1.3 - Community outreach and public relations), i.e. community workshops, and the annual international meeting, we will gather representatives from important communities, such as microbiology, virology, mycology and medical, agricultural and biotechnological societies, but also from more distant fields that could make use of the NFDI4Microbiota infrastructure and community interactions. Through these means, we aim to convey NFDI4Microbiota challenges and progress and maintain an open discussion with a range of communities to establish a widely used platform over the long term. Table 14

Table 14.
Milestones to be achieved in measure 1.6 - Sustainability.

Milestone	Month	Description
MS1.6.1	12	Initial decision-makers and funding bodies contacted
MS1.6.2	24	White paper on sustainability concept for NFDI4Microbiota
MS1.6.3	25	Options explored for further funding from federal government and states

5.1.7 M1.7 - Use cases

Contributors: RWTH (lead, 60 PM), ZB MED (co-lead, 15 PM), all co-applicants (6 PM each)

Goals: To employ all the measures offered by NFDI4Microbiota to support cutting-edge scientific projects (use cases) that involve a wide array of scientists from multiple disciplines.

This measure will help put the solutions proposed by NFDI4Microbiota into practice. All the measures converge here to serve the purpose of various use cases that represent a broad range of projects, all of which require the support of microbiota data infrastructures and services. Each use case addresses an issue of major scientific and/or global relevance that is best tackled by combining the expertise of several stakeholders acting as a small consortium with a designated lead. In all cases, NFDI4Microbiota will provide training (M1.1 - Training and education) as well as support and consulting opportunities (M1.2 - Support) to all those who need them within the community. Beyond this general support, depending on the designated goal and anticipated challenges (as mentioned in the detailed descriptions) each use case calls for different NFDI4Microbiota measures that will help overcome the main project challenges in the best possible manner.

The use cases bring together all stakeholders involved in NFDI4Microbiota, including co-applicants, participants, national research consortia (e.g. collaborative research centers, priority programs, clusters of excellence), and societies within the broad field of microbiology. Importantly, they will act as catalysts for fostering important interactions with several other NFDI consortia, in which the handling of microbiota-derived data is highly relevant but not specifically tackled.

This use case measure will not only allow the community to benefit from all the other measures implemented by NFDI4Microbiota, but will also foster improvements of these measures thanks to community feedback. Eventually, this will lead to fine-tuning of the measures over time to best address the community needs via concrete examples of scientific work performed in many labs across the country. Hence, the use cases represent a core activity within NFDI4Microbiota, for which efficient coordination is key. A lead institution has therefore been assigned to each use case, corresponding in most cases to co-applicants (since they know the consortium best and thus understand how to implement measures and coordinate actions most efficiently), but also sometimes to participants who can inject very specific expertise and carry out projects of relevance to many. This use case-specific leadership will take place under the umbrella of the measure lead at RWTH who will drive long-term commitment and efficient communication between the partners.

To implement a dynamic system that stays on track with new trends in the field, this measure will integrate new use cases from the community as they arise over time, either directly through NFDI4Microbiota activities or independently. To offer a platform for such new use cases, yearly calls for applications will be issued from year 2 onward (i.e. year 2, 3, and 4). Successful use cases will be selected via a review process organized by the

board of co-applicants and will receive financial support for a duration of 1 to 2 years via the flexible funds available within NFDI4Microbiota. Such a system will guarantee flexibility and give the opportunity to broaden the scope of activities to an even wider pool of participants, increasing the visibility and, eventually, the added-value of the consortium. Table 15

Table 15.
Milestones to be achieved in measure 1.7 - Use Cases.

Milestone	Month	Description
MS1.7.1	6	Measure policies and implementation strategy established
MS1.7.2	6	First community workshop on use cases
MS1.7.3	8	Measure strategies tailored to current needs
MS1.7.4	12	First annual report on use cases
MS1.7.5	14	Status quo of projects and dissemination
MS1.7.6	18	Annual community workshops on use cases established
MS1.7.7	20	Annual procedure established to tailor measure strategies to current needs
MS1.7.8	24	Annual use case reports established
MS1.7.9	26	Annual status quo of projects and dissemination

5.1.7.1 Use case 'AdaSPat': Phylodynamics of viral pathogens in clinical isolates

Scientific relevance: NFDI4Microbiota enables comprehensive and standardized genome-based analyses of pathogens, supporting discovery of treatment- and pathogenicity-related polymorphisms.

Case: Lilian is a clinician who has sequenced genomes of human cytomegalovirus from patient samples. She configures a tailor-made NFDI4Microbiota workflow to perform quality controls, remove human data, which falls under the GDPR, and perform viral haplotype assembly to identify mixed-strain infections. She then identifies several polymorphisms associated with changes in pathogenicity, virulence or resistances using a screen against a mutation catalog. Furthermore, the data is used together with public data to indicate putative transmission clusters.

Linked institutions and network: HZI (lead); FSU Jena; JLU; GHGA; UKJ; NFDI4Health.

Challenges and NFDI4Microbiota measures:

- Implement configurable data analysis workflows (M2.1; M2.3; M3.4).
- Interface for data visualization (M3.1; M3.2; M4.1; M4.3; M3.5).
- Manage data falling under GDPR (M2.1; M2.4; M3.5; M3.6; M4.2).

5.1.7.2 Use case ‘ATTRACTOR’: Large-scale integration of metagenomes

Scientific relevance: The number of individual microbial studies based on sequencing technologies is multiplying at a rapidly increasing rate. The research community has not yet fully benefited from this unprecedented amount of sequencing knowledge because of the lack of approaches for efficient integration and easy exploration of the data.

Case: Ilias is a young bioinformatic group leader interested in studying global diversity within microbiomes, including prokaryotes, eukaryotes, and viruses via large-scale integration of metagenomic data. In particular, he wants to prove the existence of persistent composition states (attractors) in different environments, determine the core taxa therein, and study their functional interactions. An existing public tool created during his postdoc provides a solid foundation for this project *[33]. However, he now faces the multiple challenges listed below. NFDI4Microbiota will support his work at this early stage of his academic career while also benefiting a wide variety of potential users within the community.

Linked Institutions & network: TUM (lead); RWTH; FSU Jena; EMBL; ZB-MED; SFB1371; SFB1382; SFB1076 (via FSU Jena); DataPLANT; NFDI4Earth; NFDI4Agri; NFDI4Biodiversity; VAAM.

Challenges and NFDI4Microbiota measures:

- Search, access, and curate publicly available datasets and corresponding metadata (M2.1; M2.4; M3.4).
- Rapid processing of all collected metagenomics data according to a standardized workflow (M2.3; M3.2;M3.3).
- Expand the existing interface by incorporating novel features for analyses and visualizations (M3.5;M3.8; M4.3; M3.5).
- Guarantee reliable long-term hosting of the platform (M3.6; M4.1; M4.2).

5.1.7.3 Use case ‘BIOCAT’: Biocatalysis

Scientific relevance: Novel enzymes are of central importance for making progress in biocatalysis. As well as using protein engineering methods to improve enzymes, scientists are also seeking to harness a plethora of as-yet undiscovered biocatalysts from microbes.

Case: Arthur uses high-throughput technologies to characterize novel enzymes in the Bornscheuer lab, a partner in the NFDI4Cat consortium, which already receives funding. NFDI4Microbiota will support him in studying enzymes involved in flavonoid biosynthesis such as chalcone isomerases, for which only one enzyme has been fully characterized to date. His work generates a vast amount of data that require management. Moreover, a preliminary phylogenetic analysis identified >500 candidate proteins in need of further exploration.

Linked institutions & network: Uni Greifswald (lead); RWTH; FZ Jülich; NFDI4Cat; VAAM.

Challenges and NFDI4Microbiota measures:

- Identify candidate genes/proteins from genomic and metagenomic data (M3.4; M3.7; M4.1).
- Build sophisticated phylogenetic trees (M3.2; M3.3).
- Study candidate proteins, link to biochemical and biocatalytic features, update databases (M2.1; M3.6).

5.1.7.4 Use case 'CLINIMIC': Clinical microbiomes

Scientific relevance: The number of clinical samples collected for microbial community analyses is increasing rapidly. Hence, it is essential to establish standardized operating procedures (from biobanking to data analyses) to study microbial diversity and characterize important functional traits such as antimicrobial resistances and other pathogenicity factors linked to specific clinical conditions.

Case: The University Hospital Gießen/Marburg coordinates the establishment of a nationwide biobank of bronchoalveolar lavage fluids (BALF) and paired blood samples from patients with pneumonia to identify microbial pathogens of the lower respiratory tract in combination with immune cell phenotyping by cytometry. To gain comprehensive insights into microbial communities and the virome in these samples, microbiome sequencing with different SOPs needs to be performed.

Linked institutions & network: EMBL (lead); JLU; BIBI; RWTH; UMR; FSU Jena; UK Frankfurt; UK Erlangen; UK Köln; HKI; MDC Berlin; SFB1371; SFB1382; EXC2051; GHGA; NFDI4Immuno; NFDI4Health; DGHM.

Challenges and NFDI4Microbiota measures:

- Optimize sample collection and metadata acquisition (M2.1; M2.3).
- Generate standardized microbiome profiles (M3.2; M3.5; M4.1).
- Compare with thousands of other data sets from public repositories based on securely stored clinical data (M2.4; M3.1; M3.3; M3.4; M3.8).

5.1.7.5 Use case 'CRISPR': From sequence-based diversity to microbial engineering

Scientific relevance: Microbiome research is very much hampered by the fact that only a few standard bacterial species exist as models. The vast majority of unknown commensals present in native communities represent a yet unexplored pool of functions. When cultured, commensals are often difficult to work with and manipulate, which complicates the performance of mechanistic studies.

Case: Melissa is a microbiologist who specializes in isolating and studying gut bacteria from different animal hosts by combining (meta)genomics and cultivation. Besides working on genome-based taxonomy of all her isolates and studying the sequence-based diversity of CRISPR-Cas sequences to unravel novelfunctions beyond defense, she has also identified new genes encoding lipolytic enzymes. She now wants to knock these out via phage-based delivery of target CRISPR-Cas systems and investigate their expression via

RiboSeq and quantitative proteomics, prior to studying effects on the host in gnotobiotic mice colonized with a synthetic community.

Linked institutions & network: RWTH (lead); DSMZ; UMR; JMU; FZ Jülich; SPP2330; SPP2141; NFDI4Biodiversity; VAAM.

Challenges and NFDI4Microbiota measures:

- Curation and management of generated genomes (M2.1; M2.4; M3.5; M3.6; M4.2).
- Detailed analysis of target genes (M2.3; M3.2; M3.3; M4.1).
- Development and use of molecular engineering systems (M2.2; M3.9).

5.1.7.6 Use case 'DataSci': Data Scientist

Scientific relevance: Analysis of high-throughput data obtained from various technologies requires a multitude of bioinformatic tools and databases. The NFDI4Microbiota platform will support data scientists in their daily work by providing standardized computing environments and by facilitating data management and workflow execution, regardless of the specific domain, with scalable, high-performance storage and an execution backend.

Case: Mark, an experimental biologist, has no experience with omics data analyses. He thus collaborates with Julia, a data scientist. Together they use the NFDI4Microbiota platform, which helps them manage the data and execute workflows. The platform offers a way for Mark to upload his data via a GUI or CLI, including updates and versioning, to generate reproducible results. The uploaded data is automatically analyzed via execution of a workflow created by Julia within the NFDI4Microbiota infrastructure. Julia also implemented a small web application to visualize results and deployed it via a Docker image. The versioning of all input data makes it possible to display the most recent data and compare all the data over time.

Linked institutions & network: JLU (lead); BIBI; EMBL; UK Essen; HHU; HIRI; JMU; FLI; LMU; MLU; MDC Berlin; MPI Magdeburg; TU Darmstadt; Uni Göttingen.

Challenges and NFDI4Microbiota measures:

- Train people to use the platform (M1.1).
- Build a scalable storage and computing platform with a user friendly WebUI; provide a deployment infrastructure including an authentication service (M3.1; M4.2; M4.3; M3.5).

5.1.7.7 Use case 'GUT': Gut microbiomes

Scientific relevance: As a substantial fraction of gut microbiomes is still uncharacterized, there is a dire need to implement innovative approaches to study the wealth of microbial diversity detected by shotgun metagenomics at the strain level.

Case: Jens, a microbiologist from Aachen, aims to explore the thousands of as yet unknown microbial species in the mammalian gut in order to study the host-specificity of

intestinal microbiomes at the strain level. As illustrated by the multiple challenges listed below, reaching this goal will require him to master diverse wet lab and in silico skills and process a vast amount of data.

Linked institutions & network: RWTH (lead); DSMZ; BIBI; FSU Jena; EMBL; UMR; TUM; UK Essen; HHU; UK Erlangen; UK Frankfurt; JLU; CAU Kiel; HKI; MPI Magdeburg; UK Köln; IKMB; MDC Berlin; SFB1371; SFB1382; SPP2002; EXC2051; DGHM; NFDI4Biodiversity. NFDI4Health.

Challenges and NFDI4Microbiota measures:

- Generate metagenomes and high-quality metadata from hundreds of gut samples from different host species (M2.1; M2.2; M2.4; M3.2; M3.5).
- Build metagenome-assembled genomes (MAGs) from these data and thousands of other datasets from public repositories at a strain-level resolution (M2.3; M3.1; M3.2; M3.4; M4.1).
- Establish collections of well-curated genomes, taxonomically describe and validate novel bacteria, and compare genome and MAG catalogs (M3.2; M3.3; M3.6; M4.2).

5.1.7.8 Use case ‘MetaBench’: Benchmarking meta-omics software

Scientific relevance: The plethora of tools available for metagenomics generates confusion and redundancy and enhances the risk of skewed analyses due to the lack of standards. NFDI4Microbiota will provide a platform for comparing workflows based on evaluation packages and benchmarked datasets created via the CAMI initiative (Critical Assessment of Metagenome Interpretation). This will help developers to create best-practice guidelines for common analyses and easily benchmark new techniques.

Case: Carol is a computational biologist in Braunschweig working on improving a metagenome binner developed by her lab. The benchmarking framework for meta-omics software provided by NFDI4Microbiota will help her by allowing both comprehensive testing of the binner with standard datasets and comparison to existing approaches. If Carol’s method is rated one of the best for this kind of data, it will be further analyzed internally by the platform team and eventually included in the NFDI4Microbiota data processing workflows.

Linked Institutions & Network: HZI (lead); BIBI; RWTH; EMBL; MPI Magdeburg; TUM; SFB1371; SFB1382; VAAM.

Challenges and NFDI4Microbiota measures:

- Implement the benchmarking interface (M3.1; M4.1).
- Generate further benchmark data sets (M4.3; M3.5).
- Create benchmarking workflows (M2.3).
- Derive best practices for data processing (M4.4).

5.1.7.9 Use case 'METAMAR': Global impact of climate change on marine biogeochemical cycles

Scientific relevance: Current climate change models indicate that a number of marine biomes have been transformed beyond repair. A catalog of worldwide marine microbial life (prokaryotes, eukaryotes and viruses) will not only provide a basis for comparison with future climate scenarios, but also deliver conceptual knowledge on how changes in climate may alter the functioning of biogeochemical cycles on a global scale.

Case: Jose, a marine microbiologist from Bremen, is studying the functional and phylogenomic diversity of microbes in the ocean at a global scale. He therefore decides to explore approx. 16,000 marine metagenomes publicly available in ENA. Due to the importance of multi-trophic interactions and networks in a biogeochemical context, Jose opts to examine these datasets in a multi-domain manner by studying the genetic potential and phylogenomics of prokaryotes, eukaryotes and viruses. Jose expects to create a database of >75,000 novel genomes. He will use this data to explore the biogeography of microbes and viruses, and to hypothesize potential changes of biogeochemical functioning in the marine environment.

Linked institutions and network: UFZ (lead); Uni Hamburg; FSU Jena; NFDI4Earth; NFDI4Biodiversity.

Challenges and NFDI4Microbiota measures:

- Screen public repositories for soil metagenomes and standardized metadata (M2.1; M3.1; M3.4; M3.5; M3.7).
- Long-term storage of raw data (M2.1; M4.2).
- Simultaneous recovery of prokaryotic, eukaryotic and viral metagenome-assembled genomes (MAGs) followed by phylogenomics and functional annotation (M3.2; M3.3; M4.1; M4.3; M4.4).
- Archive the novel data and associated metadata in public repositories (M3.5).

5.1.7.10 Use case 'METASOIL': Mechanisms of antimicrobial resistance and degradation of pollutants in soils

Scientific relevance: Soil microbes are the primary drivers of organic material decomposition, efflux of CO2 and other greenhouse gases, and global nutrient cycling. A better understanding of the mechanisms controlling assemblage of soil microbial communities will provide the conceptual tools necessary to understand how anthropogenic influences affect relevant microbial functions for ecosystem health (such as antimicrobial resistance and degradation of pollutants).

Case: Heike, a microbial ecologist from Leipzig, is working on the mechanisms underlying community assembly in soils systems. She therefore decides to explore approx. 8,000 datasets of soil metagenomes publicly available in ENA. Using NFDI4Microbiota metadata standardization, she will classify these different soils according to different levels of antrophogenic influences. She will then examine these communities in a multi-domain

context by studying the genetic potential and phylogenomics of prokaryotes, eukaryotes and viruses. To create a database with >50,000 novel genomes, Heike uses NFDI4Microbiota analytical tools to (i) accurately annotate the genetic potential underlying antimicrobial resistances, and (ii) define the potential multi-domain interactions involved in the metabolism of selected pollutants in soil systems.

Linked institutions and network: UFZ (lead); ZB MED; BIBI; EMBL; DataPLANT; NFDI4Earth; NFDI4Biodiversity; NFDI4Chem; NFDI4Ing; NFDI4Agri; VAAM.

Challenges and NFDI4Microbiota measures:

- Large-scale analysis of public soil metagenomes (M2.1; M3.1; M3.4; M3.7).
- Multi-domain phylogenomics and functional annotation (M3.2; M3.3; M4.1; M4.3).
- Submit novel data and associated to public repositories (M3.5; M4.2).

5.1.7.11 Use case ‘MIC-HOST’: Integration of host genetics and microbiome data

Scientific relevance: Connecting human genetics with microbiome research is scientifically powerful yet challenging due to the multiple fields of expertise and ethical considerations required.

Case: Supriya is a postdoc at University Hospital München Rechts der Isar. She is currently supervising a clinical study with 500 Crohn’s disease patients that aims to dissect the link between human genetics and gut microbial genes in inflammation. She has collected stool and blood samples and sent them for sequencing to the DFG-funded competence center in Tübingen. Patients have granted ethical approval for their data to be stored, analyzed, and integrated. She has collaborators who are experts in human genetics, Adaptive Immune Receptor Repertoires (AIRR), and microbiome analysis; Supriya herself will bring this expertise together and integrate the data.

Linked institutions: EMBL (lead); BIBI; RWTH; FSU Jena; HZI; MPI Magdeburg; TUM; UK Erlangen; Uni Hamburg; SFB1371; SFB1382; Frauenhofer CIMD; NFDI4Health; NFDI4Immuno; GHGA; DGHM.

Challenges:

- Upload patient clinical data and sample metadata in a secure and QM-compliant manner (M2.1; M3.4; M4.2).
- Integrate clinical, host genetic, and microbiome data of different formats and standards (M3.2; M3.8; M4.1).
- Synergize interactions & enable rapid data access and visualization by project partners (M3.1).
- Ensure future use of the data in compliance with safety and privacy guidelines (M2.4; M3.6).

5.1.7.12 Use case ‘MICIMG’: Microbial Imaging

Scientific relevance: Wastewater is contaminated by human activities. It contains physical and chemical pollutants and is discharged into rivers, lakes, and oceans that harbor important microbial communities. The effects of pollutants on these communities can have detrimental impacts on global ecosystem functions.

Case: Karla is a microbiologist located in Marburg. She is studying the direct effects of frequently occurring pollutants on microbial communities using microscopic imaging-based screening. Karla is especially interested in the concentration limit after which physical effects are visible (e.g. cell shape and size) within defined acquisition times. Furthermore, she wants to trace the uptake of larger waste particles and countermeasures taken by the cells including transcription profiles. Her institution is encouraged to transparently provide all data for public inspection and reuse for at least ten years.

Linked institutions and network: UMR (lead); UFZ; RWTH; JLU; IGB; NFDI4BIOIMAGE.

Challenges and NFDI4Microbiota measures:

- Determine suitable experimental setups (M2.2; M2.3).
- Convey metadata between multiple imaging experiments and omics datasets (M2.1; M3.2).
- Analyze large sets of imaging data and integrate corresponding omics datasets (M3.2; M3.8; M4.2).
- Provide long-term public access to several gigabytes of data (M2.4; M3.5; M3.6; M4.2).

5.1.7.13 Use case ‘MicroORKG’: Making knowledge from microbiological literature FAIR

Scientific relevance: As the outcome of research is primarily archived in the scientific literature, it is human-readable only and hard to query efficiently. Hence, there is a need to store statements from research articles in a systematically structured fashion.

Case: Bernd is a postdoc performing research on antibiotic resistance genes. He has compiled a large collection of scientific articles and book chapters that describe the resistance of bacterial strains from clinical isolates that show resistance against commonly used antibiotics. He now wants to share this collection and make it easily searchable.

Linked Institutions & network: ZB MED (lead); DSMZ; EMBL; JLU; TIB; NFDI4DataScience.

Challenges and NFDI4Microbiota measures:

- Research knowledge is not easily findable (M3.1, NFDI4Microbiota will host an Observatory for the Open Research Knowledge Graph *[30] and help the community to include statements; M3.3; M3.6;M4.2; M4.4).

- There is a lack of awareness of linked-open-data solutions and skills for translating knowledge into machine-readable formats in the microbiological community (M1.1; M1.2; M1.3).

5.1.7.14 Use case 'MULTI': Integration of multi-omics data of microbial species

Scientific relevance: The integration of different types of omics data such as genomics, transcriptomics, and proteomics, as well as more specific protocols such as single-cell RNAseq for in-depth analysis andChIP- and CLIP-Seq for protein-DNA and RNA-DNA interactions, can be used to detect different cellular events. This is crucial for a comprehensive understanding of the physiology of microbes.

Case: Fiona is a microbiologist in the institute of infection biology at the University of Cologne. She studies the bacterial pathogen *Pseudomonas aeruginosa* and its role in cystic fibrosis in children. Her working hypothesis is that a yet poorly described transcription factor may be involved in the regulation of infection, based on transposon insertion sequencing (Tn-seq) results obtained by another group. To test this hypothesis, she has generated a knockout mutant that she used in longitudinal infection experiments in several cell lines. She now wants to integrate multiple data types: (sc)RNA-Seq, proteomics, andChIP-Seq.

Linked Institutions & network: ZB MED (lead); JLU; UFZ; BIBI; RWTH; FSU Jena; GU; HIRI; FZJülich; JMU; KIT; FLI; LMU; MLU; MDC Berlin; MPI Bremen; RUB; UK Erlangen; Uni Greifswald;TRR124; SPP2002; VAAM.

Challenges and NFDI4Microbiota measures:

- Find related datasets with similar experimental and instrumental setups (M2.1; M3.3).
- Define requirements for each data type, including quality controls, read trimming, mapping and quantification (M2.3; M3.2).
- Integration into system biological models (M3.8).
- Consistent submission of data and connected metadata to different repositories (M2.4; M3.5; M4.2).

5.1.7.15 Use case 'PARA': DNA modifications in gametocytes of the malaria parasitePlasmodium *falciparum*

Scientific relevance: DNA methylation is an important epigenetic modification that regulates gene expression. In the gametocytes of *Plasmodium falciparum*, which are responsible for parasite transmission from the human host to the mosquito vector, DNA (de)methylation has never been studied, althoughit is likely involved in gene regulation within this deadly parasite.

Case: Jean-Pierre Musabyimana from Aachen uses Nanopore sequencing to identify modified cytosine residues in different gametocyte stages. He now intends to use a novel DNA immunoprecipitation sequencing approach (DIP-Seq) to spot specific DNA

modification sites. Whilst he feels confident with the wet lab part of the work, thanks to his training as a molecular biologist, he knows that data analysis will be a big challenge.

Linked institutions & network: RWTH (lead); ZB-MED; LMU; DGP.

Challenges and NFDI4Microbiota measures:

- Optimize DIP-Seq and implement it on experimental and clinical samples (M2.1; M2.2; M2.4).
- Analyze data from two different sequencing technologies (M2.3; M3.2; M4.1).
- Manage the data from generation to publication (M2.4; M3.5; M3.6; M3.9; M4.2).

5.1.7.16 Use case 'PATHO': Studying cohorts of bacterial pathogens isolated from hospitalacquired infections

Scientific relevance: According to the World Health Organization (WHO), antimicrobial resistance (AMR) is one of the biggest threats to global health, as it hampers prevention and treatment of an ever increasing range of infections caused by bacteria, parasites and fungi. Hence, it is essential to detect and monitor the occurrence and spread of pathogenic bacteria and their antimicrobial resistances.

Case: Helga is a medical microbiologist responsible for the routine surveillance of AMR in bacteriaand the management of nosocomial outbreaks within the university hospital in Frankfurt. Samples are routinely sequenced in the institute's local sequencing facility. However, as Helga is short of bioinformatics staff and performing data analyses is a tedious task for her, NFDI4Microbiota representsan ideal solution.

Linked institutions and network: BIBI (lead); JLU; RWTH; UK Erlangen; MDC Berlin; NFDI4Health;DGHM.

Challenges and NFDI4Microbiota measures:

- Routine sample collection and processing (M2.2).
- Structured data and metadata acquisition (M2.1; M2.4; M3.4).
- Genome processing and comparative analyses, including comparisons to publicly available genomes (M3.2; M4.1; M4.2).
- Investigation of nosocomial outbreaks and comparison with other hospital sites (M3.1; M4.4).

5.1.7.17 Use case 'PHYLOGEN': Phylogenomics of plant-associated microbiomes

Scientific relevance: Climate change is one of the biggest challenges of our times and will severely impact agriculture and food production worldwide. One important aspect to consider is how plants and their microbiomes adapt to higher temperatures, drought, and other changing environmental conditions.

Case: Christine is an environmental microbiologist studying the composition of plant species-specific bacteria under the influence of continuously increasing surface

temperature or of rising levels of atmospheric CO2 produced using free-air CO2 enrichment (FACE) technology. She needs (i) an accurate longitudinal assessment of bacterial communities based on 16S rRNA gene amplicon sequencing, and (ii) harmonized annotation, precise taxonomic classification, and phylogenomic and comparative analysis of whole genomes from selected key species. Christine is also planning to integrate her results on a higher level with data obtained from other biodiversity studies and eventually share all results with the general public to raise awareness of the impact of climate change.

Linked institutions & network: JLU (lead); DSMZ; BIBI; EMBL; UMR; GFZ; NFDI4Biodiversity;DataPLANT; NFDI4Agri; NFDI4Earth.

Challenges and NFDI4Microbiota measures:

- Routine screening of bacterial communities & quality controlled genome analysis (M2.2; M3.2; M3.5; M4.2).
- Structured data and metadata acquisition (M2.1; M2.4; M3.4).
- Robust taxonomic classification and phylogenomic analysis of fully sequenced genomes (M3.2; M3.3).
- Comparative analysis of the institute's own genomes and public genomes across different climate conditions (M3.6; M4.1).

5.1.7.18 Use case 'PLANT': Root-associated microbiomes

Scientific relevance: The rhizosphere sustains one of the most diverse host-associated microbial communities, with an estimated 30,000 bacterial species. These microbes affect plant adaptation, evolution, and health and have a real impact on agricultural sustainability. It is therefore essential to study their diversity and interactions.

Case: Rosa is a PhD student in Bielefeld working on the influence of the rhizosphere microbiome on crop plant health using shotgun metagenomics, with the additional aim of integrating metatranscriptomic and metabolomic data. In the long run, she also wants to find and compare data from other studies in the context of a meta-analysis. Hence, Rosa has lots of questions regarding optimal tools, data accessibility, and required computing power.

Linked Institutions & network: DSMZ (lead); JLU; BIBI; FSU Jena; UMR; UK Essen; HMGU; MLU;MPIPZ; TUM; Uni Hamburg; NFDI4Earth; NFDI4Agri.

Challenges and NFDI4Microbiota measures:

- Choose optimal software tools and workflows (M2.2; M2.3; M4.3; M3.5).
- Estimate and optimize computing capacities for complex analyses (M3.2; M4.1).
- Access project-related data from others with rich and standardized metadata (M2.1; M3.1; M3.3; M3.4).

5.1.7.19 Use case ‘RapSARS’: Rapid access to new SARS-CoV-2 data

Scientific relevance: Rapid processing and sharing of sequencing data is crucial in research, particularly when fast responses are required as in the case of pandemic outbreaks such as COVID-19. A standardized platform for rapid data availability could be an important breakthrough for developers of new technologies, for scientists studying the biology and epidemiology of viral infection, and for health authorities.

Case: Hortense is a clinician in charge of a large COVID-19 study looking for associations between viral genomes and clinical metadata such as disease severity. NFDI4Microbiota will offer harmonized metadata management before assigning barcodes and sticking them to the samples, which are then shipped for sequencing. Sequencing data are integrated on the fly into the NFDI4Microbiota platform, which activates automated data analysis workflows and provides electronic notifications when jobs arecompleted. In parallel, Carla, a data scientist from the health authority, has submitted a request through her user interface on the platform, asking to be alerted whenever new, relevant data becomes available. Even though she does not work directly with Hortense, she can gain access to the new sequencing data via the NFDI4Microbiota platform and explore the phylogeography of the sequenced strains and theirprevalence in Germany.

Linked Institutions & network: HZI (lead); JLU; BIBI; FSU Jena; RWTH; RKI; OvGU; ZIK Septomic;Uni Köln - IMSB; Frauenhofer CIMD; NFDI4Health; GHGA; GfV.

Challenges and NFDI4Microbiota measures:

- Implement features that allow sequencing and analysis of private and public datasets in compliance with German data privacy regulations (M2.3; M2.4; M3.4; M4.1; M4.2).
- Build synergies with NFDI4Health and GHGA to provide rapid access to sequencing data and metadata as well as search functionalities, while respecting data privacy and ethical rules (M2.1; M2.4; M3.1; M4.3).
- Implement a workflow connecting sequencing centers, clinics, scientists, and health authorities (M3.1; M4.3).

5.1.7.20 Use case ‘RNA viruses’: Pathomechanisms of RNA viruses

Scientific relevance: Virus infections remain a major threat to human health. RNA viruses are of particular interest because their replication machinery introduces a high number of nucleotide substitutions, leading to enhanced adaptive behaviors. A recent example is the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), which caused the current COVID-19 pandemic.

Case: Lisa is a physician working in pulmonary research. She is currently investigating the pathomechanisms of virus-induced lung injury, especially in patients infected with SARS-CoV-2. Her methods include analyzing the transcriptome of both host cells and the virus in time-series in vitro experiments by combining bulk and single-cell RNA sequencing. This

results in large, complex datasets, for which multiple bioinformatic pipelines are required, which are out of her area of expertise.

Linked institutions and network: FZU Jena (lead); JLU; HHU; SFB0221; Frauenhofer CIMD;NFDI4Health; GfV.

Challenges and NFDI4Microbiota measures:

- Standardize experimental design for bulk and single-cell RNA-seq (M2.2).
- Routinely collect clinical samples of virus-infected host cells (M2.4).
- Structured data and metadata acquisition (M2.1; M3.4; M4.2).
- Standardize data processing & comparative analyses (M3.2; M3.8).
- Combine analysis of data from several hospital sites and epidemiology (M3.1; M4.4).

5.1.7.21 Use case 'SMALLPRO': Prokaryotic small PROTEINS in gut metagenomes

Scientific relevance: Less than half of all proteins produced by gut microbes can be annotated, not to mention those with false annotations. Specialized small proteins encoded by Biosynthetic Gene Clusters (BGCs) represent an important, understudied group of molecules within the gut microbiome, since their biosynthesis is usually a multi-step process and they may have relevant functions such as anti-microbial activities. It is therefore important to explore gut metagenomes to provide detailed characterization of such molecules.

Case: Jaymi, a PhD student in Aachen, has established multiple collections of bacterial strains from the intestines of humans and other animals. In collaboration with TU Dresden and SPP2002, she identified a potentially new family of BGC-encoded small proteins in the genome of some of her strains. NFDI4Microbiota will be able to help her take this project further, e.g. by screening the genomes of all other strains for additional, similar discoveries and by performing targeted investigations of the small proteins she has already identified.

Linked institutions & network: RWTH (lead); EMBL; HKI; SPP2002; SFB1371; SFB1382;NFDI4Biodiversity; VAAM, DGHM.

Challenges and NFDI4Microbiota measures:

- Genome sequence processing and curation (M2.1; M2.4; M3.2; M3.5; M3.6; M4.1; M4.2).
- Quantitative measurement of identified small proteins by proteomics (M2.2; M2.3; M3.9).
- Impact of the target small molecules on the gut microbiome in vivo by multi-omics (M2.2; M3.8).

5.1.7.22 Use case 'Soft-DEV': Scientific software developer

Scientific relevance: Contemporary bioinformatics software development often requires the integration of various public and private data repositories. Thus, the storage platform of

NFDI4Microbiota will support scientific software developers by providing access to data and metadata via highly standardized frameworks.

Case: Lev is a software developer who designed a new algorithm which requires the analysis of metadata. He can access the metadata stored along with biological data using the standardized REST API provided by NFDI4Microbiota and further process this metadata using the well-documented and easily accessible JSON format. Further, large-scale analyses are supported by the storage platform's high-performance backend. Finally, Lev is able to directly integrate the REST interface into his new applications, allowing users to easily access the data. An authentication layer can be used to restrict access and upload to the platform using ELIXIR AAI.

Linked institutions & network: JLU (lead); BIBI; EMBL; FSU Jena; GFZ; FLI; RSE4NFDI.

Challenges and NFDI4Microbiota measures:

- Train people to use the platform (M1.1).
- Build a scalable storage platform (M3.1; M4.2; M4.3; M3.5).

5.1.7.23 Use case 'SPACE': Microbial isolates from outer space

Scientific relevance: Public and private space organizations are eager to resume crewed space travel targeting the moon and Mars. Investigating microbes within the closed habitat of spacecraft during longer space journeys is important, as exposure to harsh space conditions (high levels of UV and ionizing radiation, space vacuum, low gravitational force) may drive harmful mutations.

Case: Juri is a microbiologist at the DLR studying bacterial isolates obtained from space missions. He is responsible for monitoring bacterial and viral contamination based on swipe samples from the ISS and hardware assembly facilities of the European Space Agency (ESA). These samples, and the single microbes they contain, are routinely sequenced and need to be analyzed in a standardized manner, forwhich Juri urgently needs support.

Linked institutions & network: DLR Cologne (lead); JLU; FSU Jena; UMR; GFZ; PUNCH4NFDI.

Challenges and NFDI4Microbiota measures:

- Routine collection of microbial samples and preparation for sequencing (M2.2; M2.3).
- Structured data and metadata acquisition (M2.1; M2.4; M3.3; M3.5).
- Genome processing and comparative analyses with third-party data (M3.2; M3.4; M3.6; M4.1; M4.2).

5.1.7.24 Use case ‘Strain-ID’: Collecting and matching microbial strain identifiers

Scientific relevance: Strain identifiers represent a universal language for scientists, clinicians and industry, as well as for legal compliance and regulatory control purposes. They enable users to generate and communicate findings and to compare and verify data, ensuring traceability and reproducibility. Whilst unique identifier systems exist (INSDC accessions, E. C. enzyme numbers), microbes are referenced only via numbers issued by culture collections and informal historical identifiers.

Case: André is a postdoc in mycology who inherited a strain collection of biotechnological interest from a retired colleague. Unfortunately, due to inconsistencies in strain identification in the literature, it is difficult for him to appreciate the value of his strains. He remembers that the database *straininfo.net* may be useful, but discovers it has been put permanently out of commission. With >700 culture collections worldwide using different collection numbers, the task of tracing and obtaining reference strains will be a laborious one.

Linked institutions & network: DSMZ (lead); RWTH; Uni Göttingen; TRR124; NFDI4Biodiversity;DGfM.

Challenges and NFDI4Microbiota measures:

- Build a database that collects and matches strain identifiers from different sources and provides stable referencing as well as linking with other identifiers, e.g. INSDC sequence accession numbers (M2.1; M2.4; M3.1; M3.3; M4.3; M3.5).

5.1.7.25 Use case ‘SYNSYS’: Synthetic systems

Scientific relevance: Polystyrene is an extremely recalcitrant pollutant accumulating as litter in the environment. Developing sustainable and scalable methods to degrade this pollutant is of utmost global relevance.

Case: Philipp, a synthetic biologist from Marburg, works with Denise, a microbial ecologist from Leipzig, on microbe-based degradation of polystyrene litter. They aim to implement a synthetic pathway for digestion that is scalable to large-scale industrial use. Their collaborative project will include both laboratory and computational work and will be carried out at two separate locations with different backgrounds but entangled objectives. Proper harmonization and sharing of data is therefore essential.

Linked institutions and network: UMR (lead); UFZ; ZB-MED; JLU; HMGU; FZ Jülich; IGB; MPIMagdeburg; TU Darmstadt; NFDI4Cat; NFDI4Chem; NFDI4Ing; VAAM.

Challenges and NFDI4Microbiota measures:

- Screen public omics and text data for relevant enzymes for pathway construction (M2.1; M3.1; M3.4; M3.7).
- Determine adequate workflows for genetic construction, testing and optimization (M2.2; M2.3).

- Evaluate and integrate omics and imaging data including datasets from third parties (M3.2; M3.8).
- Manage, organize, and share large amounts of data (M2.4; M3.5; M3.6; M4.2).

Overall schedule, contribution, risks, and dependencies

Fig. 8

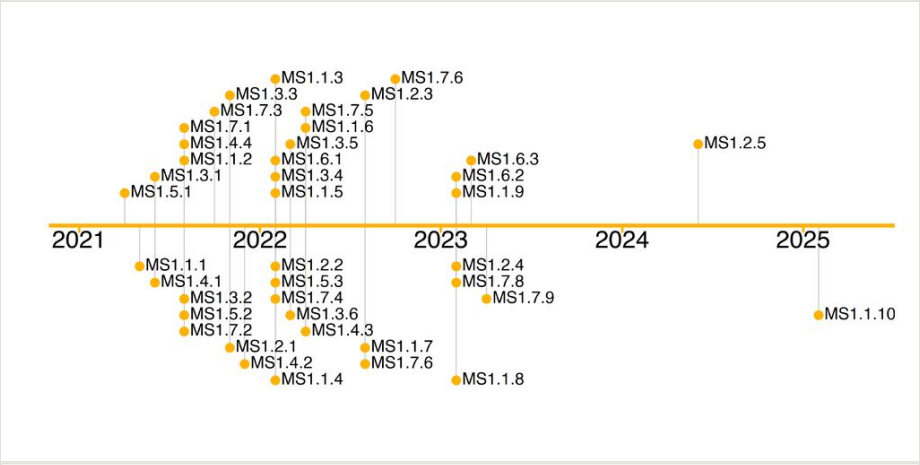

Figure 8. doi

Overall schedule Task Area 1.

Contributions: Of all the task areas, TA1 brings together the highest number of contributions by all the co-applicants, just ahead of TA3 on services. This clearly demonstrates how – aside from the technical aspects – NFDI4Microbiota places a great deal of importance on activities linking this consortium to the community and beyond via training, support, outreach, and collaboration. Contributions by the lead and co-lead (HZI and ZB-MED) are highest, followed by RWTH (all three institutions with more than 100PM) for coordination of the use cases, which are key instruments for bringing all stakeholders together around collaborative scientific projects. The contribution of all other co-applicants averages around 60PM/institution. Together, the co-applicants provide extremely solid foundations for this TA through their in-kind contributions – in the form of a multitude of existing training and support activities – and through their networking with multiple participants, other NFDI consortia, and international partners. Forthcoming contributions within NFDI4Microbiota will include coordinating and harmonizing existing measures to boost them further, as well as developing new measures to stay at the forefront of a rapidly moving field of research.

Dependencies: The success of TA1 is intrinsically linked to the entire landscape of microbiota research in Germany, since NFDI4Microbiota will act as an important interface between the co-applicants, who bring together a rich portfolio of expertise, and a community that represents a multitude of needs. Building bridges between the consortium

and this community is the biggest dependency of TA1, but also its biggest strength, since the consortium will benefit from community feedback in return. This task area also relies on all the other TAs and measures implemented by NFDI4Microbiota, since training, support, and use cases are catalysts of community interaction based on all the various tools, measures, and services developed by the consortium. This broad scope of dependencies is, in fact, a strength, as progress within TA1 will not be hampered by unexpected events that may slow down the progress of one specific arm of the technical working program.

Risks of implementation and mitigation approaches: (1) There is a risk that the target communities will not be strongly engaged by NFDI4Microbiota. However, we think this risk is low, because we developed the concepts for this infrastructure in close collaboration with the community and consistently received very positive feedback. Continuing to keep the community actively involved is one of the consortium's most important goals, as illustrated by its dedicated training and outreach plans. Should this not suffice, we will use our flexible funds (M5.3) to invest further in community outreach and engagement and consider their feedback even more extensively in all consortium activities, as well as further leveraging our own contacts within these communities. (2) Another risk is that the demand for training could exceed available capacity since, in our experience, the demand for training in computational methodologies is usually very high, with double or triple overbooking. Should this happen, we will allocate more flexible funds to training, prioritize training courses and participant selection inconsultation with the SAB, and increase our focus on creating reusable, scalable online training resources and making use of existing high-quality national and international training resources. (3) Further, it may prove difficult to identify adequate strategies for guaranteeing sustainability and long-term maintenance, as this is a major challenge for all infrastructures. However, the consortium members are all very experienced and well-connected, so we believe suitable long-term solutions will be found. We also firmly believe that the extended training we are planning will help make this infrastructure a success and guarantee its long-term usage.

5.2 Task area 2 - Standards & Policies

Coordinators: EMBL (lead), DSMZ (co-lead)

Building on the work of others is a pillar of scientific research that greatly enhances the results of public investment. However, it can be challenging to make use of existing data and scientific findings due to inadequacies in the recording and sharing of (meta)data and the sheer variety of methodologies used, some of which may be poorly documented. Having consistent, easily accessible, documented standards is a prerequisite for effectively and efficiently reusing and integrating data. NFDI4Microbiota will address these challenges by facilitating and stimulating the sharing and reuse of data. It will achieve this by supporting easy, open access to microbial and associated data in line with FAIR principles. The measures in this TA will promote the use of community-recognized and expert-curated data and metadata requirements (M2.1 - Data & metadata standards), experimental protocols (M2.2 - Experimental procedure standards), and analytical approaches (M2.3 - Workflow standards). Continuous community input and discussion of these standards will

also be supported in order to build connections between researchers. This resource of standards and protocols will be a core of the NFDI4Microbiota system and will support the creation, recording, and analysis of microbialdata that is standardized, repeatable, and reusable. These activities will be guided by policies that encourage collaborative research, facilitate user compliance, ensure high-quality data, and stimulate integration of public data (M2.4 - Policies and legal issues for data management and reuse). These policies and the systems they guide will ensure open access, reproducibility, consistency, transparency, and interoperability throughout NFDI4Microbiota's services.

Fig. 9

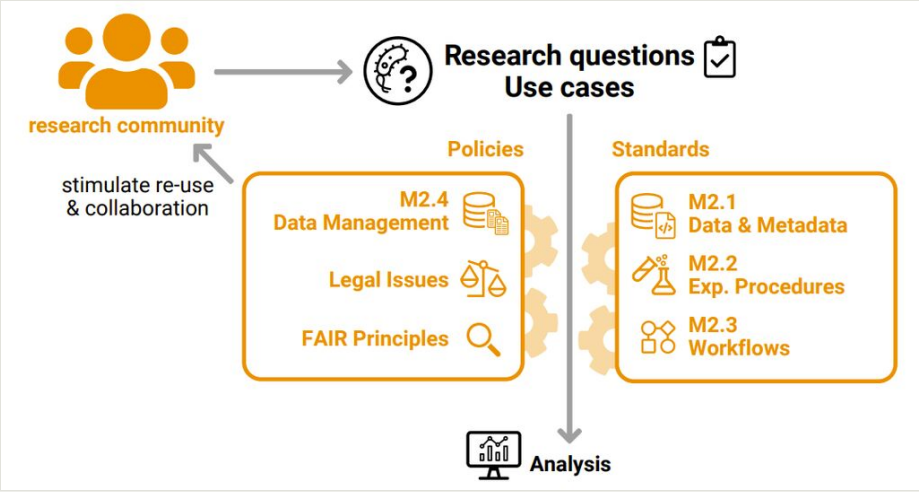

Figure 9. doi

Overview of measures in Task Area 2.

5.2.1 M2.1 - Data & metadata standards

Contributors: UFZ (lead, 69 PM), FSU Jena (co-lead, 19 PM), DSMZ (6 PM), EMBL (4 PM), UMR (4 PM), RWTH (8 PM), JLU (6 PM)

Goals: To maximize the quality of data entering the NFDI4Microbiota system by enforcing compliance with existing standards, as well as to identify and promote additional tailored data standards and metadata requirements within the NFDI4Microbiota systems.

Pipelines that use data from omics or state-of-the-art microbiota analysis rely on the quality of the data and the completeness of the metadata. The diversity of data types, how they are produced and the different metrics used to assert their quality are not always clear, particularly for non-experienced users. As data quality is of the utmost importance for any subsequent analysis, this task area will enforce compliance with existing standards and, where necessary, define further standards for the various techniques relevant to NFDI4Microbiota in collaboration with international initiatives (M1.5 - Connection to

international partners). We will use the commenting/reviewing services (M3.7 - Reviewing/commenting system for data) to improve the interface used for data quality for different domains of data. We will also request feedback from scientists at different career levels and adjust which types of data are relevant. The standards employed by the NFDI4Microbiota consortium will ensure that only high-quality raw data entered together with sufficient and standardized metadata is analyzed by the NFDI4Microbiota resources. Further, standardized metadata will facilitate future integration and comparison of the data with novel state-of-the-art analyses or new data. Measure M2.1 will define open, easy-to-use interfaces to check if raw data and metadata meet the quality standards desired by the NFDI4Microbiota Networks and will provide users with access to these interfaces. To define the standards for raw data quality, we will use the capacity of different co-applicants and participants specialized in the different types of data. Metadata standards will be categorized as technical (i.e. dependent on the technology used for the data), biological (i.e. dependent on the microbiota analyzed) or environmental (i.e. defining the environmental variables relevant for those samples). The technical metadata will be defined by the various data-type specialists within the NFDI4Microbiota Network while the biological and environmental metadata will be defined by specialists from various fields of the consortium (e.g. human microbiomes, terrestrial microbiome and marine). We will enforce compliance with standards defined by key initiatives such as: the Genomic Standards Consortium (Field et al. 2008), which defines the minimum necessary information on an omic sample (Yilmaz et al. 2011), which contextualizes biological and biomedical entities; and the OBO Foundry (ref), which coordinates the evolution of ontologies to support biomedical data integration. The NFDI4Microbiota team has already worked on metadata standardization for metagenomes from terrestrial environments (Corrêa et al. 2020). To avoid the bottleneck of users having to search for all the information required for their specific data and metadata once the standards for data and metadata are defined, we will generate user-friendly applications to enable users to check the quality of their data and metadata. Therefore, Measure M2.1 will link data generation and downstream analysis by creating materials that train users how to measure the quality of the various data types supported by the NFDI4Microbiota consortium. The development and definition of metadata standards will be carried out in close cooperation with M4.2 - Data storage platform to ensure efficient metadata handling in the storage backend. Table 16

Table 16.

Milestones to be achieved in measure 2.1.

Milestone	Month	Description
MS2.1.1	4	Data standards defined for the different types of raw data
MS2.1.2	8	Data standards defined for technical metadata
MS2.1.3	16	Online system launched to check data quality
MS2.1.4	22	Online system launched to check metadata standards

Milestone	Month	Description
MS2.1.5	24	Annual procedure established to revise and add standards for technical and biological metadata
MS2.1.6	36	Online tutorials established on data and metadata standards
MS2.1.7	48	Online automated system established to check raw data quality
MS2.1.8	60	Online automated system established to check metadata quality

5.2.2 M2.2 - Experimental procedure standards

Contributors: EMBL (lead, 41 PM), DSMZ (co-lead, 3 PM + sig. in-kind), RWTH (12 PM), FSU Jena (12 PM)

Goals: To increase the uniformity and interoperability of data collected within the NFDI4Microbiota network by identifying and promoting best practices for performing microbial research.

A major source of biological variation in experimental results comes from differences in how experiments are performed and how biological materials are collected and processed. While statistical methods exist to counteract these barriers to data integration, and will be made available to users through M3.2 - Analytical services, these approaches are not always sufficient to overcome data incompatibilities. The centralization of information in the NFDI4Microbiota network provides the opportunity to streamline and unify experimental procedures from the start, which would improve future meta-analysis outcomes (M3.8 - Systems biology, modeling, and multi-omics integration), prevent duplication of effort in developing SOPs, and spread expertise between fields (M1.4 - Connection to other NFDI consortia). **This measure will provide users with an open, version-tracked resource of curated standard operating procedures (SOPs) for experiment design, sample collection, storage, and processing in culture-based and microbiome studies.** These SOPs will provide guidance at both a general level (e.g. sampling of any microbial community) and for specific types of starting material (e.g. sampling microbial communities from human faeces). SOPs will be collected from NFDI4Microbiota participants, existing initiatives (e.g. IMHS (www.microbiome-standards.org/), the Earth Microbiome Project (https://earthmicrobiome.org/), the European Microbial DNA Bank Network (www.microdnabank.eu/), and peer-reviewed publications. The NFDI4Microbiota SOP resource will improve on existing resources by curating and integrating SOPs from distinct experimental approaches to encourage the production of integratable data types. Initial SOPs selected will be for well-established experimental approaches. As best practices evolve and new experimental techniques mature, SOPs will be added and updated, guided by the NFDI4Microbiota Ambassador Council. SOPs will be provided to the community in the centralized NFDI4Microbiota web system M3.1 - Central web portal. This will support commenting, which will facilitate community dialogue around the SOPs, revisions, which will ensure SOPs can evolve with new best practices, and version-tracking, which will support referencing of exact SOPs, thereby facilitating reproducible research. Table 17

Table 17.
Milestones to be achieved in measure 2.2.

Milestone	Month	Description
MS2.2.1	12	SOPs collected, centralized and curated
MS2.2.2	18	Training materials and corresponding web interface developed to guide correct usage of experimental SOPs
MS2.2.3	20	Community-accessible online system launched for commenting, updating, and versioning of SOPs
MS2.2.4	36	Annual SOP update established, including addition of newly collected SOPs to online resource, report on user-suggested revisions, and incorporation of selected revisions into main system
MS2.2.5	44	Training materials updated
MS2.2.6	60	Annual update of SOPs and training materials established

5.2.3 M2.3 - Workflow standards

Contributors: HZI (lead, 60 PM), BIBI (co-lead, 27 PM), JLU (4 PM), RWTH (8 PM), EMBL (3 PM),FSU Jena (12 PM)

Goals: To develop guidelines for standardized and FLOSS-compliant workflows, and to establish standard workflows for data quality control and benchmarking services. To create maximum value for the community, NFDI4Microbiota will offer standardized workflows for users to analyze their own datasets as well as running these on curated datasets created and provided by the consortium. Work in this task area will focus on developing best practices for standardized workflow design and development, as well as for microbiome data processing (e.g. using Common Workflow Language (CWL), Snakemake, Nextflow). The consortium will ensure that (i) workflows are developed according to FLOSS principles (Free/Libre/Open Source Software), (ii) code is accessible, (iii) software is versioned with stable releases, (iv) software is containerized (e.g. using Docker) to ensure reproducibility of results across different computing environments, which otherwise might differ, and (v) the application programming interfaces (APIs) are made public. In addition, data provenance will be tracked within the workflows. We will establish joint working groups with national and international partners (e.g. EMBLEBI MGnify, EOSC *[24], the ELIXIR Interoperability Platform and de.NBI) to develop international standards in workflow design, e.g. emphasizing a modular approach to enable exchange of workflow parts. We will also implement strategies for workflow maintenance and performance optimization (e.g. runtime and memory consumption) and ensure sustainability and high quality standards for implemented workflows by emphasizing documentation and automatic testing. The EDAM ontology (Ison et al. 2013) will be used to define types of data and data identifiers, data formats, operations and topics for the established workflows. EDAM provides a set of concepts with preferred terms and synonyms, definitions, and additional information.

To complement quality control performed by data stewards, we will establish standardized quality control workflows for data including sequencing data, data processing results and metadata, such that all data held by the NFDI4Microbiota platform complies with the data and metadata standards established in M2.1 - Data & metadata standards. The created workflows and results will be strictly versioned according to the principles of semantic versioning, indicating minor (e.g. version 1.0.1 to 1.0.2) and major software changes (e.g. version 1.1.0 to 1.2.0). Data generated with different releases of the same software will be labeled accordingly, reinforcing comprehensibility and reproducibility of findings based on data processed with the NFDI4Microbiota platform. We will also develop a process for integrating user-defined, standard-compliant workflows into our system (together with M4.4 - Service monitoring & reporting). These will undergo a quality review process before being offered as an option to the community.

The computational services offered, i.e. the functionalities available within workflows, will be based on user requirements (e.g. as determined in M1.3 - Community outreach and public relations) and bestpractice recommendations determined in systematic benchmarking efforts, such as ELIXIRs OpenEBench platform and CAMI, the Initiative for the Critical Assessment of Metagenome Interpretation (Sczyrba et al. 2017), organized by A.C. McHardy and A. Sczyrba. The infrastructure will thus offer users the most suitable approach for each task, so that they can perform analyses of their data without needing to be experts in the plethora of computational methods used in that area. Implementation of specific functionalities will be prioritized by expected user demand and community relevance.

Where indicated, we will benchmark methods for individual use cases (e.g. workflows for depositing and analyzing clinical microbiome datasets in interaction with GHGA, M1.7 - Use cases). To this end, we will provide benchmarking options and a representative collection of metagenome benchmark datasets, enabling users to assess software for computational microbiome research. This service will build on community-derived benchmarking concepts developed e.g. in CAMI and ELIXIR, such as metrics and visualizations embedded in benchmarking software, Docker-based standardized virtualizations of tool categories (e.g. metagenome assemblers) in bioboxes (Belmann et al. 2015) and datasets (Meyer et al. 2021, Sczyrba et al. 2017). This will facilitate the comparison of developed methods with existing approaches and the development of performanceoptimized analytical workflows.

Communication of developed standards and best practices will be connected to (M1.3 - Community outreach and public relations; M1.1 - Training and education). Table 18

5.2.4 M2.4 - Policies and legal issues for data management and reuse

Contributors: UMR (lead, 48 PM), JLU (co-lead, 2 PM), RWTH (12 PM), DSMZ (6 PM)

Goals: To develop and implement policies for data management, sharing and reuse that ensure quality and openness of data but also legal certainty. Our policies will respect the interests of users, provide guidance on making responsible use of available resources, and

safeguard compliance with the FAIR principles, thus stimulating the reuse of data and fostering new research.

Table 18.
Milestones to be achieved in measure 2.3.

Milestone	Month	Description
MS2.3.1	12	Best practices developed for workflow implementations and benchmarking service
MS2.3.2	12	Standards developed for reproducible data analysis workflows and automated data quality control
MS2.3.3	18	Prototype best-practice workflow implementation established
MS2.3.4	24	Prototype benchmarking service established
MS2.3.5	30	Best-practice workflow implementation finalized
MS2.3.6	40	Benchmarking service productive
MS2.3.7	24	Annual best-practice data processing recommendation releases established

NFDI4Microbiota offers a broad range of guidelines, integrated tools, and workflows for the preparation, processing, and analysis of research data (M2.1 - Data & metadata standards, M2.2 - Experimental procedure standards, M2.3 - Workflow standards, M3.2 - Analytical services, M3.8 - Systems biology, modeling, and multi-omics integration). This data needs to be evaluated, stored, organized, secured, backed up, shared, and, where necessary, safely removed. For this purpose, data management plans will be created and managed using the Research Data Management Organizer (RDMO) (Neuroth and Engelhardt 2018) tool. These plans will also cover all intermediate files, result files, and visualizations that are generated in the process of an analysis. The management plans will be translated into applicable policies that are then implemented in the workflows and, wherever possible, applied automatically within implemented procedures and workflows. Compliance with policies will be supervised and monitored regularly by internal implementation reviews and functional real-life tests where actual data will be used that might violate current policies. The initial stage will be a concept of management plans and policies that follow internationally established recommendations and guidelines on data sharing e.g. by the Research Data Alliance (RDA) (RDA COVID-19 Working Group 2020). Based on this concept, we will acquire feedback from potential users, participants, and members to compile a set of definitions that is generally agreed upon within the consortium. This process will be iterated several times during the progression of NFDI4Microbiota in order to cover the latest needs and developments, to harmonize with other NFDI consortia such as NFDI4Biodiversity, DataPLANT, NFDI4Health, NFDI4Earth (M1.4 - Connection to other NFDI consortia) and international partners (M1.5 - Connection to international partners), and to maintain interoperability with major data providers such as the centers of the Next Generation Sequencing Competence Network (NGS-CN), EBI, NCBI and UniProt, whose usage we explicitly foster.

NFDI4Microbiota will develop a framework of operational principles and guidelines that ensures open access, reproducibility, consistency, transparency, and interoperability throughout the whole data workflow. Compliance with FAIR principles will be supported by developing all platform components as free and open source software (FOSS) within our NFDI4Microbiota consortium. This includes the application of best-practice guidelines for software engineering and community interaction, as well as the use of OSI-compliant licenses. To foster collaboration with other NFDI consortia, we will ensure that all development processes are transparent to outside users and software developers. Generally applicable guidelines ensuring open access, reproducibility, consistency, transparency, and interoperability will be embedded in the training (M1.1 - Training and education).

Legal issues are an overarching topic that needs to be addressed as a cross-topic by the whole NFDI (Leipzig-Berlin declaration (Bierwirth et al. 2020)). Nevertheless, microbial research poses specific challenges that need to be considered from the start (e.g. how to handle remnant/trace amounts of host genomic information, risk classification). NFDI4Microbiota will therefore be at the forefront of efforts to educate and support microbial researchers. This includes the incorporation of legal issues and how these should be handled in training (M1.1 - Training and education).

The legal conformity of data and metadata will be ascertained upon submission. NFDI4Microbiota will ensure that all work performed within the consortium complies with the legislation in force that governs the use of personal health information data, including the World Medical Association (WMA) Declaration of Helsinki and the WMA Declaration of Taipei. In addition, the General Data Protection Regulation (GDPR, EU-R 670/2016) will be considered in light of additional national or state-specific regulations. Conditions for processing and sharing by NFDI4Microbiota will be sufficiently anonymized and, if applicable, additional data filtering will be applied (e.g. removal of contaminant human sequence fragments from shotgun metagenomes). The TA leaders will coordinate the structures and processes of NFDI4Microbiota with regard to ethical and privacy issues in close cooperation with the Technologie und Methodenplattform für die vernetzte medizinische Forschung e.V. (TMF) and other NFDI consortia such as GHGA and NFDI4Health (M1.4 - Connection to other NFDI consortia). Legal conformity will be re-assessed regularly in case new methods arise that could re-enable the traceability of individuals.

The *Nagoya Protocol* *[34] is a legally-binding international instrument that came into force on October 12, 2014 and established procedures for access and benefit-sharing. For scientists working internationally – even within different EU member states – this means that there are new legal requirements in place when accessing (sampling) or using biological material. Although Germany is a so-called free access country, which means there are no *Nagoya* obligations on German material, German scientists are required to follow the laws of other countries and once *Nagoya*-relevant documents have been obtained, these need to stay connected to the organisms/material.

The implemented policies will be strict enough to ensure legal conformity, high-quality data and adherence to the FAIR principles (Wilkinson et al. 2016), but this should not prevent researchers from using the platform. User acceptance is an equally crucial aspect that must be addressed. A typical outcome will be that data is not publicly accessible to anyone except explicitly mentioned collaborators before the research question is answered or the respective publication is finished. Even if it seemingly contradicts the idea of open science, temporary data locks and limited access to sensitive data will play a necessary part in making data accessible in the first place. Similarly, our data management policies will ensure the data produced by our users has a long life cycle, making it easily findable and widely accessible and helping to foster new research. Table 19

Table 19.
Milestones to be achieved in measure 2.4.

Milestone	Month	Description
MS2.4.1	6	Initial concept established for data management plans and polices
MS2.4.2	9	Review of legal issues and release of training documents established
MS2.4.3	15	Policy set agreed on by the consortium performed
MS2.4.4	18	Monitoring concept established for implemented policies
MS2.4.5	28	Documentation and implementation of data management policies established, training documents released
MS2.4.6	32	Documentation and implementation of data anonymization and filtering policies established
MS2.4.7	40	Documentation and implementation of sharing and reuse policies established, training documents released
MS2.4.8	24	Annual reiteration of concepts, harmonization with other consortia established
MS2.4.9	31	Annual monitoring and testing of implemented policies established

Overall schedule, contribution, risks, and dependencies

Fig. 10

Contributions: Reflecting the importance of TA2 and the expertise that all co-applicants bring, 9/10 co-applicants will contribute to this TA. Further, participants will be encouraged to contribute to and comment on some standards and policies. Major leadership roles will be taken on by UFZ for data and metadata (with support from BIBI), EMBL for experimental procedures and overall coordination (with support from DSMZ), HZI for workflows, and UMR for data management and reuse. Several institutes will support most measures, facilitating their connectivity (DSMZ, EMBL, FSU, JLU, RWTH).

Dependencies: The creation and curation of standards and of an online resource to disseminate them to the microbiological community will be supported by the in-depth

microbiological, ethical and research expertise of the co-applicants and of the participants. Establishing standards is a moving target, as standards may evolve over time as the field evolves. This dependency on ongoing developments is mitigated by the strong ties that NFDI4Microbiota has to the multiple stakeholders in the field, which will ensure it is made aware of any changes that may need to be made to standards. The development and use of the online resource will depend on the web interface (M3.1), and, importantly, on training, support, and raising awareness among users (M1.1, M1.2, M1.3).

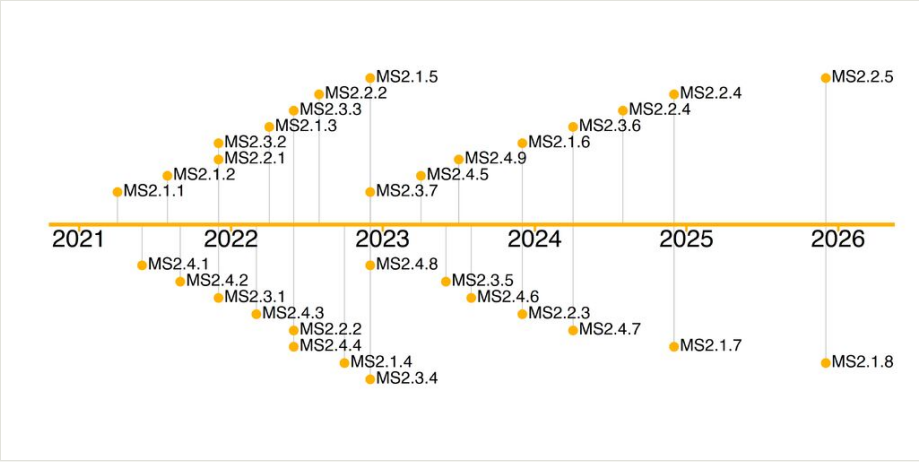

Figure 10. doi

Overall schedule Task Area 2.

Risks of implementation and mitigation approaches: The major risk to the success of this Task Area concerns the uptake of and adherence to these standards and policies, which will depend on the engagement of users. The risk of user disuse or misuse will be minimized through provision of comprehensive training, support and outreach across Germany (M1.1, M1.2, M1.3). This will occur both through the organized, official activities of NFDI4Microbiota, as well as through community relationships, which are significant given the wide coverage of institutes and fields among NFDI4Microbiota's coapplicants and participants, and the emphasis placed on community building (M1.3). A secondary risk involves disagreements arising in the selection or promotion of some standards over others. This will be mitigated by taking an inclusive approach to the presentation of standards and supporting a user forum in which benefits and limitations can be openly discussed. The use of respectful and inclusive communication will be enforced.

5.3 Task area 3 - Services

Coordinators: JLU-Gießen (lead), FSU Jena (co-lead)

One central aim of the NFDI4Microbiota consortium is to offer services for data access and analysis to the microbiome community that enable standardized and reproducible research. This task area is dedicated to achieving this goal within the NFDI4Microbiota

platform. To this end, NFDI4Microbiota will realize and operate a central web portal, the NFDI4Microbiota Hub. This will offer centralized access to the corresponding services (i.e. various analytical services), a means of collecting analysis results, the ability to access data and search across multiple databases and services, information on training events, and support (M3.1 - Central web portal). The platform will provide analytical workflows to store, access, process, and interpret the various data types generated in microbiome research. A range of workflows for common microbiome data processing demands will be developed and maintained, based on state-of-the-art software components and visualizations (M3.2 - Analytical services). To encourage structured and lasting data accessibility, support will be provided for database development and best practices, including, for example, development of a strain identifier database (M3.3 - Databases & terminology services). Stored data will undergo vigorous quality control and well-documented data provenance will be ensured to lay the groundwork for interpretable, comparable and valid research results (M3.4 - Data quality & provenance services). To foster usage of the platform and cultivate the quality of uploaded data, the NFDI4Microbiota interface will be user-friendly and measures will be taken to promote best practices for archiving data in repositories (M3.5 - Data deposition & repositories). Because a central aspect of any infrastructure is ensuring its long-term usage, NFDI4Microbiota will offer and promote the standardized digital preservation of metadata, as well as raising awareness of the importance of this topic (M3.6 - Long-term preservation). To this end, the consortium will also develop a means for users to review and comment on NFDI4Microbiota services and analyze the feedback to enable user-prioritized changes and improvements to the platform (M3.7 - Reviewing/commenting system for data). As integration with complementary data is important for any cross-cutting research task, NFDI4Microbiota will develop mechanisms for mapping linkages between data types supported by NFDI4Microbiota and inter-omic integration between reference databases and the NFDI4Microbiota system (M3.8 - Systems biology, modeling, and multi-omics integration). Finally, NFDI4Microbiota will promote and educate on the usage of Electronic Lab Notebook (ELN), because these are an invaluable tool for providing research data and experiment documentation, and thus for reproducible and synergistic research (M3.9 - Electronic Lab Notebooks).

Fig. 11

5.3.1 M3.1 - Central web portal

Contributors: ZB MED (lead, 60 PM), EMBL (co-lead, 18 PM), DSMZ (6 PM), JLU (4 PM), UFZ (12 PM)

Goals: To provide numerous and wide-ranging services to the microbiology community in Germany and beyond. These will be offered through a one-stop shop, a central web-based portal known as the **NFDI4Microbiota Hub**.

Integration of services: The portal will act as an information resource while also giving direct access to the services NFDI4Microbiota provides to the community. For this purpose, it will integrate the various internal analytical services (M3.2 - Analytical services), collect

analysis results, link to external resources, provide access to data, and enable users to search for datasets and query across multiple databases and services (M3.3 - Databases & terminology services). It will also provide connections to tools and resources not (yet) included in the NFDI4Microbiota platform by linking to existing service listings, such as bio.tools *[35].

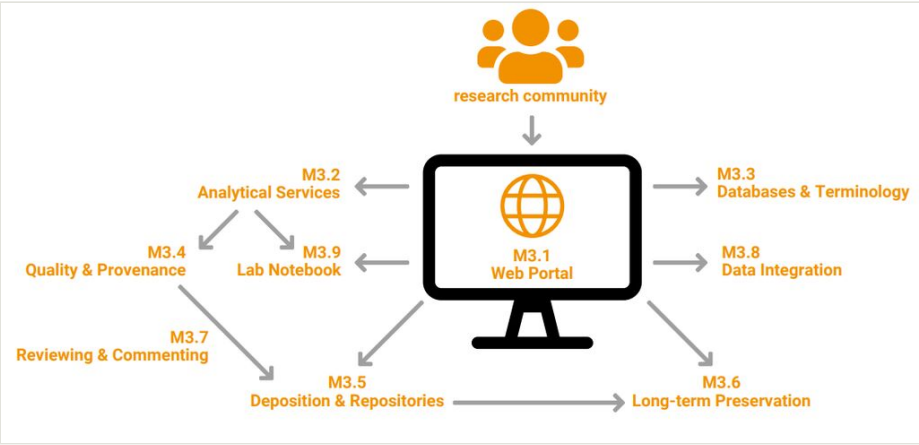

Figure 11. doi

Overview of measures in Task Area 3.

Furthermore, it will allow users to sign up for training events (M1.1 - Training and education), provide access to help desks (M1.2 - Support), and serve as a marketplace to bring together researchers with complementary expertise and list job vacancies.

Subportals: Having a single, central portal will create an anchor point for the diverse communities brought together by NFDI4Microbiota's activities. However, to facilitate usage of domain-specific tools and avoid confronting users with overwhelming options, we will also create tailored portals adapted to the needs of NFDI4Microbiota sub-communities, such as COVID-19 researchers, cell culture specialists, and bioengineers. These tailored portals will maintain clear ties to the central portal through their similar look and feel and consistent structure. Once users are comfortable using the website within their domain of expertise, they can easily move on to other areas of the website.

Search: The ability to search for datasets efficiently is crucial. The NFDI4Microbiota Hub will include a powerful discovery service that will also enable users to search based on biological and/or experimental setups. Literature search functions will be integrated through ZB MED's discovery service LIVIVO *[36] with its underlying ZB MED Knowledge Environment and state-of-the-art semantic search technologies.

API: Besides the web interface, the portal will also provide a rich API for querying metadata and performing automated bulk uploads and downloads. It will feature a user-friendly front-end for data submission (building on the data deposition back-end developed in M3.5 -

Data deposition & repositories) and allow users to curate metadata directly within the web interface. Table 20

Table 20.
Milestones to be achieved in measure M3.1 - Central web portal.

Milestone	Month	Description
MS3.1.1	6	Basic web portal launched
MS3.1.2	12	Inclusion of pilot services completed
MS3.1.3	15	Help desk inclusion completed
MS3.1.4	24	REST-API established for queries and bulk uploads/downloads
MS3.1.5	36	Cross-platform meta search launched
MS3.1.6	50	Data submission and curation interface finalized
MS3.1.7	60	Sub-community-specific portals established

5.3.2 M3.2 - Analytical services

Contributors: HZI (lead, 90 PM), BIBI (co-lead, 27 PM), JLU (18 PM), EMBL (16 PM), UMR (5 PM + sig. in-kind), FSU Jena (21 PM), RWTH (12 PM)

Goals: To develop and maintain best-practice analytical workflows and services for processing microbiological research data.

Fundamental to the aims of the NFDI4Microbiota consortium is the development and provision of the computational infrastructure and analytical workflows required to store, access, process, and interpret various microbiome-related data types. This data may be generated by a range of use cases (M1.7 - Use cases) and may include omics data of microbial communities, bacterial and archaeal isolates, as well as fungi, parasites and viruses (see M3.3 - Databases & terminology services). To develop workflows, NFDI4Microbiota will establish synergies through international cooperation, making use of reproducible, open source solutions wherever possible (e.g. *[24]), and interacting, for instance, with working groups of the ELIXIR Interoperability Platform and teams working on EMBL-EBI's MGNify database, JGI IMG Integrated Microbial Genomes & Microbiomes, MG-RAST, and the Qiita platform (Gonzalez et al. 2018) (R. Knight, UCSD). We will also carry out methodological benchmarking in community efforts such as CAMI (M2.3 - Workflow standards). Workflows, query interfaces and visualization implementations will be prioritized based on estimated user numbers and range of hosted data types. Required methodologies include software for quality control, data processing, statistical analyses, and visualizations of different data types and results.

Innovations and missing software components will be developed and integrated by the NFDI4Microbiota partners as needed, following good software engineering practice and

according to the FLOSS principles (Free/Libre/Open Source Software). Further high-quality workflows developed by users will be allowed to run within the framework, providing they meet the standards established in M2.3 - Workflow standards. All developed software will be made public under Open Source Initiative (OSI)-compliant licenses. NFDI4Microbiota will cooperate with the RSE4NFDI for this purpose.

NFDI4Microbiota will also work with international partners to establish mechanisms for sharing, integrating and combining user-relevant workflow outputs such as metagenomic assemblies and MAGs, and depositing them in ENA. To ensure a high-quality user experience and excellent support, we will work together with M1.2 - Support on establishing a help desk and a user forum. The user forum will enable users to support and interact with each other and serve as a community-based knowledge resource. Additionally, we will work with M1.1 - Training and education to offer dedicated training courses on the workflows we develop, since we believe that training is key to guaranteeing active use of the NFDI4Microbiota infrastructure and ensuring its success. Tables 21, 22

Table 21.

Preliminary list of key services that could be offered through the NFDI4Microbiota platform as curated, high-quality workflows of state-of-the-art software.

Data type	Preliminary list of services by data type
Metagenome	Functional profiling, taxonomic profiling, assembly, MAG recovery, differential profile analyses, metabolic reconstructions, biomarker discovery, identification of key genes, pathogen discovery
Marker gene data (16S, ITS, etc.)	Taxonomic profiles, community diversity (alpha, beta-diversity), biomarker discovery
Metatranscriptome	Differential expression, functional and taxonomic profiles, metatranscriptome assembly, metabolic pathway analyses
Metaproteome	Differential protein abundance, mechanism of action discovery (e.g. to investigate how a bacterium responds to a drug). Functional and taxonomic profiles (also in combination with metagenomics and/or metatranscriptomics data)
Microbial, fungal, viral genomics	Genome assembly and annotation, functional profiles, metabolic reconstructions, biomarker discovery, evolutionary analyses, comparative genomics, orthology prediction, phylogenomics
Microbial, fungal, viral transcriptomics	Differential expression, metabolic analyses
Microbial proteomics	Functional profiling, mechanism of action discovery, differential protein abundance
Metabolomics	Differential metabolite abundance, metabolite profiling
Microscopic imaging	Cell counting, cell shape / morphology determination, species distinction, quantification of fluorescence, protein-protein interactions, cell tracking, particle tracking

Table 22.
Milestones to be achieved in measure M3.2 - Analytical services.

Milestone	Month	Description
MS3.2.1	12	Concept developed for data quality evaluation/ preprocessing toolbox and for initial set of workflows
MS3.2.2	13	Workflow and tool training curriculum defined together with M1.1 - Training and education
MS3.2.3	15	Toolbox for data quality evaluation and processing finalized
MS3.2.4	25	Workflow prototypes implemented, process in place for continuous updating
MS3.2.5	30	Workflows functional on platform back-end
MS3.2.6	37	Initial release of workflows, process in place for continuous updating

5.3.3 M3.3 - Databases & terminology services

Contributors: DSMZ (lead, 120 PM), FSU Jena (co-lead, 60 PM), RWTH (10 PM), UMR (2 PM), EMBL (4 PM), ZB MED (14 PM), UFZ (sig. in-kind)

Goals: To support FAIRification in database development, safeguard knowledge in databases, and build up central databases for collecting and matching strain identifiers as well as for non-pathogenic virus data.

Support FAIRification in database development and safeguard knowledge in databases: Efficient access to comprehensive and standardized data is essential for large-scale analysis. Modern highthroughput methodologies are generating data at ever-increasing rates. Tremendous amounts of data are being collected, yet there is a risk that much of it will never be reused. To make this data accessible, databases have been developed to collect and standardize scientific data. Except for a few centrally managed databases, most of these lack long-term funding. The majority of databases are developed as part of a project or a PhD thesis and are orphaned when the scientist responsible leaves. Although many of these databases are valuable, they often become inactive and are ultimately lost.

NFDI4Microbiota will establish a help desk to support researchers during all phases of a database's life cycle (M1.2 - Support). This will include advising researchers during development, supporting them during database operation, and ascertaining when a database is no longer being actively developed. During development, many aspects of standardization and interoperability can be optimized, greatly enhancing the value for the scientific community.

Content of orphaned databases needs to be preserved, e.g. by relocating databases to a centrally managed system. In the simplest scenario, data is transferred to a long-term archive as a data dump, described with rich metadata, and thus made findable for researchers. A more sophisticated approach is to transform the data and transfer it to a

public database (e.g. Wikidata) where researchers can directly access and reuse the data. The ultimate goal is to successively transfer data into a machineinterpretable format such as the Open Research Knowledge Graph (www.orkg.org) to improve access and enable semantic queries.

To ensure and enhance data findability, NFDI4Microbiota will assist researchers by providing advice (M1.2 - Support), offering training (M1.1 - Training and education), and carrying out active development to make sure that all data sources are available through the **central web portal** (M3.1 - Central web portal).

Central database for collecting and matching identifiers for prokaryotic and eukaryotic strains: Terminology services are of great importance for the NFDI as a whole, since unified, consistent identification and understanding of entities provides the basis for FAIR data. In the case of microbial data, terminologies are partially covered by the NFDI4Biodiversity terminology services, and NFDI4Microbiota will support an NFDI-wide terminology service. Still, there are gaps in the terminology of microbial research that need to be addressed.

One very significant gap is the lack of unique and stable identifiers for microbial strains, which are a prerequisite for the comparability and reproducibility of microbial data. The lack thereof therefore represents **a major obstacle to FAIR microbial research** (Use case 'Strain-ID': Collecting and matching microbial strain identifiers).

NFDI4Microbiota will build a database to collect and match strain identifiers from different sources. This will offer (1) a one-stop-shop for strain identifiers searchable via the **central web portal** (M3.1 - Central web portal) and (2) a Web service (API) for the automated retrieval of strain identifier information. (3) Constant efforts will be made to curate and match strain identifiers as well as to explore and integrate new sources including strain identifiers representing non-culturable strains. Up-to-date information will be published every six months. (4) Moreover, this service will be closely coordinated with relevant activities in other NFDIs, especially NFDI4Biodiversity.

Central virus database: Viruses are of utmost importance for NFDI4Microbiota (M1.7 - Use cases, e.g. Use case 'GUT': Gut microbiomes and Use case 'METASOIL': Mechanisms of antimicrobial resistance and degradation of pollutants in soils). The storage of virus genome data is essential and is currently a computationally unsolved problem. In future, we will need novel, qualitatively different computational methods, possibly pan-genomes to reflect the virus quasispecies as a cloud of viral haplotypes. Currently, besides the general NCBI database, a few virus-specific databases exist, such as the Virus Pathogen Resource (ViPR) *[40] with a restricted focus on human pathogens, GISAID EpiFlu for Influenza A with restricted access to data, HIV DB, HCV DB, and Viralzone as the comprehensive encyclopedia for viruses without sequence storage. The surprising lack of a comprehensive viral database is the rationale behind this measure. NFDI4Microbiota is invited to use the data structure, established tools, and resources of ViPR to build a complementary virus database for non-human viruses and non-pathogenic viruses (M1.5 - Connection to international partners). The two databases will share the same underlying

structure and tools, and will update each other regularly without holding content in several places and while adhering to data and metadata standards (M2.1 - Data & metadata standards). EMBL-EBI is planned as the storage location for the central virus database. This database will benefit substantially from a (non-trivial) community acceptance with an established high service grade, including several tools and web portals (M3.1 - Central web portal). Additionally, usage training and education will be offered (M1.1 - Training and education). Table 23

Table 23.
Milestones to be achieved in measure M3.3 - Databases & terminology services.

Milestone	Month	Description
MS3.3.1	18	Strain identifier information collected and curated; new database built for strain identifiers; updates published every 6 months
MS3.3.2	30	API developed for automated retrieval of strain identifier information; strain identifier integrated into central web portal and other NFDI services, e.g. NFDI4Biodiversity terminology service
MS3.3.3	18	Central virus database acquired and adapted as extension of ViPR; updates published every 6 months
MS3.3.4	24	Suitable data structure released for viral MAGs

5.3.4 M3.4 - Data quality & provenance services

Contributors: UMR (lead, 44 PM), DSMZ (co-lead, 3 PM), UFZ (sig. in-kind)

Goals: To foster qualified and interpretable research by ensuring data quality and well-documented data provenance.

Data quality is a major prerequisite for qualified, interpretable, and valid research results * [37]. Another often underrated aspect is data provenance, which allows the tracing of data sources and processing steps. Well-documented provenance helps to confirm the authenticity of data and allows researchers to assess the comparability of studies. These aspects become even more pressing in the light of open science, which promotes reproducibility, reuse, and sharing of data. Ensuring full transparency by monitoring quality and keeping track of provenance is a resource-heavy and time-consuming task that, moreover, requires knowledge of best practices. This can become a significant obstacle to research and research groups. NFDI4Microbiota will therefore provide intuitive services to address data quality and provenance throughout the processing and analysis workflows.

Based on the type of data (M2.1 - Data & metadata standards) and the chosen workflow (M2.3 - Workflow standards), data quality will be evaluated automatically when the data is initially uploaded and after each processing step. A brief quality summary will be stored in the metadata to ensure full transparency. The detailed report can be reproduced at any time as per our consistency and reproducibility policy (M2.4 - Policies and legal issues for data management and reuse). Reporting will focus, in particular, on significant problems.

Typically, users are left to cope with quality warnings on their own. In contrast, our platform will give them access to expert experience and enable them to take advantage of defined experimental procedures (M2.2 - Experimental procedure standards) and workflows (M2.3 - Workflow standards & M3.2 - Analytical services). The interpretation of quality warnings will be supported by the clarification of potential implications for the interpretation of any results and best practices. Where possible, expert advice will be given on how to proceed or step back and improve the data. This “feedback assistance” offer will be extended regularly based on frequent use cases and requests. Furthermore, the results of feedback and advice will be integrated in the respective training (M1.1 - Training and education). Most quality checks will be mandatory to ensure compliance with our quality standards.

Provenance data will be kept as metadata in the form of a knowledge graph. We will set up a framework that transfers and extends metadata along the process of processing and analyzing datasets. Processing steps, filters, tool versions and parameters, quality checks, and so on will be documented automatically without requiring any user interaction. If a user decides to modify a processing step due to the outcome of quality control, this will be noted and reasoned in the metadata as well. When using large workflow pipelines, provenance data can easily become complex and hard to follow. Thus, our service also includes a means of inspecting provenance data in the form of graph-based visualizations. Moreover, this data will be used to automatically generate a “materials and methods” text for the given dataset which will describe the processing and analysis steps in a human-readable form and greatly facilitate subsequent manuscript preparation. This text will include references to tools, databases and data sources, thereby ensuring seamless documentation of the work and supporting FAIR citation practices. Table 24

Table 24.

Milestones to be achieved in measure M3.4 - Data quality & provenance services.

Milestone	Month	Descriptiom
MS3.4.1	4	Concept established for provenance tracking & data quality checks
MS3.4.2	12	Collection and integration of expert advice established
MS3.4.3	16	Provenance tracking implemented
MS3.4.4	28	Data quality checks implemented
MS3.4.5	32	Feedback assistance implemented
MS3.4.6	40	Human-readable export of provenance data implemented
MS3.4.7	50	Release of training material based on collected cases

5.3.5 M3.5 - Data deposition & repositories

Contributors: JLU (lead, 30 PM), UFZ (co-lead, 20 PM), ZB MED (12 PM)

Goals: To develop and provide standardized interfaces and an extensible platform for submitting data to, and searching for data in, public repositories.

Data repositories such as the European Nucleotide Archive (ENA) are a valuable resource for facilitating standardized deposition and reuse of generated data. In addition, these repositories can be used as an off-site backup platform and to provide widely recognized digital object identifiers (DOIs). Submitting data to these repositories, however, is a time-consuming and non-trivial task. In many cases, this hinders researchers from depositing their data into such a repository even after their study has finished, especially if their project did not require such a submission.

To facilitate the process of data submission, this measure will provide a platform to easily submit data to a set of well-recognized repositories using a REST API. This API can be used to directly submit data from our storage system, which is developed in M4.2 - Data storage platform, to an appropriate repository. The system will handle all the repository-specific aspects of the submission and check if the data and metadata is correctly formatted and the metadata contains all necessary information for a submission. The API can be accessed either directly using ELIXIR AAI, or via the central web portal developed in M3.1 - Central web portal. Depending on the user's preference, the data can also be deposited in private mode until publication, if supported by the corresponding repository. We will also implement these mechanisms for repositories that only allow submission through a website. Depending on the type and volume of data, we will also support an optional direct submission of the data to a public repository. It is additionally planned to automatically propose suitable repositories for data deposition based on the provided metadata.

Besides submitting data to public repositories, another important part of data handling is the ability to access and search for data in public repositories. We will develop a mechanism to find and load data from public repositories. Users will be able to perform queries across multiple repositories simultaneously. Frequently accessed repositories and datasets can be manually or automatically cached inside our storage infrastructure to avoid unnecessary repeated transfer of the same datasets. We will combine our efforts with NFDI4Biodiversity, who are planning a similar mechanism, to develop this system, if the requirements match.

Certain types of datasets, especially those that contain highly sensitive data, require special care before submission to external repositories. Thus, we will collaborate closely with NFDI4Health and, in particular, GHGA to make these datasets accessible under appropriate conditions of data security and privacy. This will include assisting researchers in handling these datasets correctly prior to submission.

The implementation of these features will be based on an extensible architecture. The wrapper implementations for the individual repositories will follow object-oriented programming design principles to build a collection of separate modules (repository adapters), which are embedded into the platform. This modular approach allows us to easily integrate modules that are developed in other communities as well, since this

consortium alone will probably not have sufficient resources to implement adapters for all relevant public repositories. During the initial development, we will select 2-3 highly relevant repositories to showcase and test the service components and to collect user feedback as quickly as possible. Table 25

Table 25.
Milestones to be achieved in measure M3.5 - Data deposition & repositories.

Milestone	Month	Description
MS3.5.1	9	Stable API and beta testing of data deposition component
MS3.5.2	12	Stable API and beta testing of data query component
MS3.5.3	14	First stable release of data deposition component
MS3.5.4	20	First stable release of data query component
MS3.5.5	24	Annual integration of additional data submission interfaces for public data repositories
MS3.5.6	36	Annual integration of additional data query interfaces for public data repositories

5.3.6 M3.6 - Long-term preservation

Contributors: ZB MED (lead, 12 PM), UFZ (co-lead, 5 PM), DSMZ (in-kind)

Goals: To establish digital preservation solutions for data types that do not have dedicated repositories. This includes enriching data with metadata, developing standards, raising awareness, and providing training and consulting on this topic for community members.

Providing digital preservation services: NFDI4Microbiota will provide the microbiology community with (1) bitstream preservation of (meta)data in the de.NBI cloud (M4.2 - Data storage platform), (2) preparation of files (e.g. formatting) for digital preservation, and (3) digital preservation of (3.1) molecular data in EMBL-EBI repositories and of (3.2) other data types in ZB MED's dark archive through publication in the PUBLISSO Repository for Life Sciences (FRL) *[38]. The FRL enables authors to publish and archive life sciences research data in a way that makes them permanently accessible and citable. This data may be linked to a publication or self-contained (i.e. raw research data). Research data published in the FRL must be licensed as open data (i.e. opening up the possibility of subsequent reuse) and provided with a detailed description and metadata to ensure it can be clearly interpreted and reused in the future. ZB MED intends to make it possible to publish open-source software on the FRL. The FRL is based on Rosetta, a scalable digital preservation system built upon the OAIS reference model and relevant metadata standards such as Dublin Core, PREMIS and METS (M2.1 - Data & metadata standards). Rosetta is run as a "dark archive" to protect the data against unauthorized access. The digital resources available in the FRL can be searched via ZB MED's LIVIVO discovery service *[36]. The specific search functionality for objects from the microbiological community will be implemented in the NFDI4Microbiota Hub (M3.1 - Central web portal).

Fostering standardization and quality assurance: To standardize and control the quality of archived (meta)data, NFDI4Microbiota will select and implement relevant metadata standards (e.g. PREMIS). Furthermore, it will select suitable file formats for digital preservation (M3.5 - Data deposition & repositories) and advise data producers and repository operators on file format and integrity checks (e.g. format identification, validation, checksums). Additionally, the consortium will foster collaboration between research data producers and other repository operators on topics such as technology and community monitoring, and standardization.

Awareness building and training: To make digital preservation an indispensable part of microbiological research projects, NFDI4Microbiota will train researchers in the digital preservation of research data and associated metadata, for instance during Research Data Management workshops (already established by ZB MED, M1.1 - Training and education). Furthermore, the consortium will partner with the research community to help define criteria for assessing the archival value of research datasets. We will also offer individual support to researchers (M1.2 - Support) to assist them with specific issues related to digital preservation. Table 26

Table 26.
Milestones to be achieved in measure M3.6 - Long-term preservation.

Milestone	Month	Description
MS3.6.1	6	Dedicated training material generated
MS3.6.2	12	SOPs established for data publishing and annual reviewing process
MS3.6.3	12	SOPs established for digital preservation and annual reviewing process
MS3.6.4	48	Publication of open-source software on FRL enabled

5.3.7 M3.7 - Reviewing/commenting system for data

Contributors: UFZ (lead, 12 PM), DSMZ (co-lead, sig. in-kind)

Goals: To create and promote an online platform for community development of NFDI4Microbiota systems using reviews and comments from users.

Reviews and comments are the medium that connects developers (in this case the NFDI4Microbiota consortium) and users. Reviews and comments are also the way users communicate with each other. In the NFDI4Microbiota network, they will be used not only to ensure that the different NFDI4Microbiota systems are connected to their users, but also as a form of input that allows the community to participate in the development of the various systems. Measure M3.7 will provide an online platform for the NFDI4Microbiota network that will enable users to review and comment on all the different systems for data. The online reviewing and commenting platform will be built on the basis of five key principles: 1) Build a connection between NFDI4Microbiota users and developers. We will use our reviewing/commenting platform to encourage users to ask questions and give

suggestions. By doing so, we will give users the chance to comment freely on the different systems they use. This will also create a direct link between users and the developers of the different systems and help to make the systems more successful. 2) Create connections among NFDI4Microbiota users. We will create a platform where users can answer questions raised by other users. This will add a community-driven dimension to the development of the NFDI4Microbiota systems. 3) Develop a medium for improving user interaction. User discussions will be searchable, providing a long-term platform for community development and forging connections between users. 4) Create backlinks. Backlinks are links from one website to a page on another website. These play an important role in most search engine algorithms. The traffic created by reviews and comments will link the various NFDI4Microbiota systems to the keywords from the respective user reviews and comments. This will make NFDI4Microbiota visible not only to other NFDIs but also to anyone searching for keywords relevant to the NFDI4Microbiota Network. 5) Users can rank the relevance of reviews and comments from other users. This will be used to help NFDI4Microbiota system developers rate the relevance of questions and comments and prioritize changes and new developments in the system. In other words, reviews and comments will be used as tools to make the NFDI4Microbiota system more community-driven. Table 27

Table 27.
Milestones to be achieved in measure M3.7 - Reviewing/commenting system for data.

Milestone	Month	Description
MS3.7.1	2	Contact established to the developers of the different NFDI4Microbiota systems
MS3.7.2	4	Development of reviewing and commenting template complete
MS3.7.3	9	Links created between reviewing and commenting platforms following the launch of the various systems
MS3.7.4	12	Annual process established to analyze comments and reviews in order to identify which changes and developments to prioritize within the NFDI4Microbiota systems

5.3.8 M3.8 - Systems biology, modeling, and multi-omics integration

Contributors: EMBL (lead, 46.5 PM), UFZ (co-lead, 23 PM), ZB MED (sig. in-kind)

Goals: To make it easier for users to contextualize their data and extend the scope and depth of their research through facilitated integration with complementary data.

Data integration can greatly enhance the value of individual datasets by increasing the magnitude and diversity of sampled populations and the biological facets they measure. The typical workflow for molecular profiling of microbial cultures or communities involves quantitative screening of one or more data types – such as 16S RNA profiling, (meta)genomics, (meta)transcriptomics, (meta)proteomics, and metabolomics – to provide a systems-level view of microbiota and microbial communities, enable meta-analyses, and

inform modeling of biological processes. NFDI4Microbiota will facilitate integration between different data types within a study (i.e. the same set of samples measured in different ways) and between data from different studies (different sets of samples, measured in the same or different ways). Integration will have three goals: (1) to provide linkages from a user's dataset to relevant public datasets and databases that add context and knowledge, (2) to make it easier for users to combine and compare their data with public data in order to add context and boost confidence in their results, and (3) to enhance the quality of the analyses that can be performed on a dataset and expand the biological questions that it can answer through multi-omics integration. This measure will design and implement integrative analyses and make them available to users through the central web portal and/or the analysis interface (M3.1 - Central web portal and M3.2 - Analytical services).

Integrative data linkage: User access to public complementary data and relevant biological models will be supported though integrative data linkage, which will help users discover connections between studies and with relevant databases. For example, genetic data from an in vitro microbial drug exposure experiment could be linked to metagenomic data from human fecal samples from patients who took these drugs, allowing the user to determine whether in vitro toxicity is reflected in decreased species abundance in situ. This will be possible thanks to the consistent way in which researchers upload their metadata connected to an experiment (M2.1 - Data & metadata standards). More generally, existing public data that are relevant to a user's dataset will be computationally identified and gathered from the NFDI4Microbiota system (enabled by M3.1 - Central web portal and M3.3 - Databases & terminology services) and reported centrally to the user (enabled by M3.2 - Analytical services).

This integrative data interface will support inter- and intra-omic linkage between datasets based on the structured data generated from Task Areas 2 and 3 and the storage infrastructure in M4.2 - Data storage platform. Some of these linkages will then further support integrated analyses across datasets.

Integration of analysis results: When analyzing a dataset, users often want to compare their results to those of previous similar studies, or run a meta-analysis across studies. This can be challenging if the data from other studies are not easily accessible, the metadata is not properly linked to the data, or the analysis tools are computationally expensive or difficult to run. The NFDI4Microbiota platform provides a valuable opportunity to easily combine and compare results from many datasets. This is possible because the exact software versions and running parameters will be known for all results generated from analyses executed on the NFDI4Microbiota platform (supported in M3.2 - Analytical services). The results of consistent analyses run on public datasets can therefore be programmatically gathered and merged with a user's results, providing valuable external context or validation. Datasets will be selectable thanks to their well-formatted and consistent metadata (M2.1 - Data & metadata standards, M3.5 - Data deposition & repositories).

Synergistic data analysis: The combined analysis of different profiling technologies – such as 16S RNA profiling, (meta)genomics, (meta)transcriptomics, (meta)proteomics, and

metabolomics – can be synergistic and often adds confidence and value to molecular readouts. Combining different omics techniques can further provide a multifaceted view of a biological system and improve the creation of biological models of cellular processes in microbial species, as well as of their interactions with biotic and abiotic factors, the latter being fundamental for understanding microbial functionality and being predictive at an ecosystems level. Combining different types of data for microbial communities is, however, a very challenging task which requires a high level of expertise in, and access to, different technologies and is therefore lacking in many studies. This integration will thus be guided by benchmark studies and achieved through integrative analyses that will help microbial researchers to perform and develop cutting-edge multi-omics research. Multiple omics data types will be integrated using existing and newly developed tools and approaches. The biologically informed integration of different data types leads to new biological questions and thus improves the quality of research results.

By providing a unified, centralized database of diverse microbiological data with consistent storage, processing, and metadata formats, the NFDI4Microbiota platform creates the opportunity for powerful data reuse and integration. In this measure, this potential is harnessed to create synergy across datasets and experimental approaches and empower systems-level microbiological research. Table 28

Table 28.

Milestones to be achieved in measure M3.8 - Systems biology, modeling, and multi-omics integration.

Milestone	Month	Description
MS3.8.1	8	Mapping completed of linkages supported by NFDI4Microbiota infrastructure between and within data types and to external databases
MS3.8.2	18	Workflows developed to integrate analysis results
MS3.8.3	26	Analysis services adapted to handle multi-dataset input
MS3.8.4	32	Generation of links between user's analysis results and complementary data (analysis results and/or reference databases) established
MS3.8.5	40	Profile-based integration of analysis results from (multi)-omics data established
MS3.8.6	48	Machine learning pipelines and online tutorials developed to select multi-omics features predictive of metadata categories (bioindicators)
MS3.8.7	50	Deployment of workflows for integrated analysis across datasets completed
MS3.8.8	54	Release of training materials for integrative analysis services
MS3.8.9	60	Workflows adapted based on user feedback

5.3.9 M3.9 - Electronic Lab Notebooks

Contributors: ZB MED (lead, 12 PM), RWTH (co-lead, 16 PM), FSU Jena (in-kind)

Goals: To foster the use of a common Electronic Lab Notebook (ELN) framework and create guidelines for selecting and using ELNs.

ELNs play an important role in documenting research data: they provide clear documentation of experiment planning and implementation and of data generation and processing. The aim of this measure is to ensure research data are captured as early in the process as possible in order to feed them directly into the analysis pipeline.

ELNs provide a digital means of organizing all the necessary metadata, which can then be used to create or update data management plans (DMPs). Over half of those who responded to our community questionnaire say that they intend to use, or are already using, an ELN solution. Additionally, the results indicated that more than 10% of respondents were not aware of ELNs as a potential solution. This shows a clear need to build awareness.

This measure seeks to cover two cases: firstly, situations where ELNs are already in use and, secondly, situations where help is required in choosing an appropriate tool. In the first case, various different ELN systems are available; we therefore intend to provide a range of interfaces to ensure compatibility, especially since there are currently no data exchange standards for ELN tools. In the second case, the selection of a suitable ELN can be aided by a needs analysis in the co-applicant's or participant's institution. It is important to remember that a variety of users will work with the ELN, including researchers, laboratory specialists, laboratory managers, and IT administrators. Each user will have their own specific requirements. Some requirements are common to all users, including the ability to map the most important lab-specific workflows and processes and fully document the research process. Criteria for the needs analysis will be based on the ELN Guide, which aims to help users select an appropriate ELN tool *[39]. Key criteria include usability, compliance with good scientific practice, the option of integrating the tool in institutional research data management workflows, IT security, and costs. The output of a needs analysis mainly comprises answers to the following questions: Is there a need for one or more ELNs? Do existing commercial or open source ELN tools match the users' needs, or is it necessary to develop a more personalized solution?

Since ELNs can be used in just about any field of research, we intend to work on this topic in close cooperation with other consortia. By pooling resources with partners in the life sciences and beyond (e.g. NFDI4Chem), we can take the development of open source software to the next level. The key is to compile the needs of the microbiology research community in a clearly structured way. Table 29

Table 29.

Milestones to be achieved in measure M3.9 - Electronic Lab Notebooks.

Milestone	Month	Description
MS3.9.1	3	Overview created of ELN tools used by co-applicants and participants (via survey)
MS3.9.2	9	Provision of interfaces established for data exchange between ELNs that are already in use

Milestone	Month	Description
MS3.9.3	12	Needs analysis completed to aid in the selection of one or more ELNs
MS3.9.4	18	Work plan established for developing or purchasing one or more ELNs
MS3.9.5	24	Annual community data exchange workshop established

Overall schedule, contribution, risks, and dependencies

Fig. 12

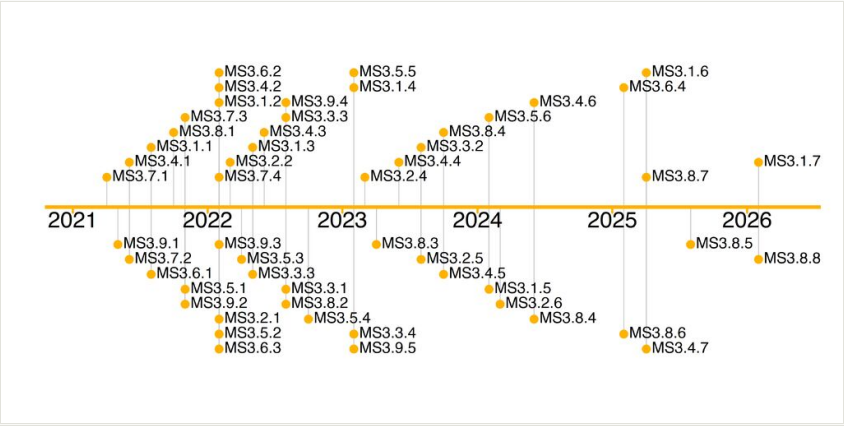

Figure 12. doi

Overall schedule Task Area 3.

Contributions: The majority of work in this task area evolves around the provisioning of services for end users. All co-applicants are involved in at least one of the individual measures. Major leading roles are taken by ZB MED, HZI, DSMZ, UMR, JLU, UFZ, and EMBL. Each co-applicant will contribute their broad knowledge to the development of the various components and services and apply their experience to creating user-friendly applications. This includes competence and experience on a technical level as well as domain-specific knowledge in the various domains of microbial research.

Dependencies: The services and workflows implemented in TA3 will utilize the infrastructure provided by TA4. This includes use of the various platforms developed in M4.3 for the provisioning of DBMS as well as workflow execution backends such as SLURM or Spark clusters for more advanced scalable data analysis. It also includes use of the storage platform developed in M4.2 for storing data objects along with their associated metadata. In addition, the individual components will be monitored using the tooling developed in M4.4. All metadata handling, which is a cross-cutting topic within all parts of TA3, is based on the metadata standards defined by M2.1 and the metadata validation system developed in M4.2. The workflow standards from M2.3 are required for M3.2 and M3.4 to implement a proper provenance service.

Risks of implementation and mitigation approaches: In the event that the planned coordination between the measures of TA3 and TA4 does not prove to be efficient enough, the involvement of software architects will be increased in each of the relevant measures. Moreover, in the event of any inconsistencies or a failure to adopt standards from M2.1 and M2.3, we will boost the overlap and exchange of personnel and agree on more detailed requirement specifications. Finally, if the usability of the offered platforms turns out to be sub-optimal, we will use the feedback from user surveys to tailor the platforms to community needs.

5.4 Task area 4 - Technical Infrastructure

Coordinators: BIBI (lead), JLU-Gießen (co-lead)

This task area is dedicated to providing the technical foundations for the NFDI4Microbiota platform. A central goal of NFDI4Microbiota is to offer compute and storage resources to its community.

To this end, TA4 will explore and define resource requirements for the planned NFDI4Microbiota services, especially for services provided by TA3, but also for the NFDI4Microbiota community in general. By leveraging synergies to existing infrastructures such as the de.NBI Cloud, TA4 will provide and maintain a suitable, scalable, high-performance, cloud-based compute infrastructure (M4.1) as well as a platform for scalable and accessible data storage (M4.2). Common computational frameworks for analytical and storage services, as required either by other measures (e.g. Analytical Services M3.2) or by tool developers from the NFDI4Microbiota community, will be provided in M4.3. To ensure and foster community acceptance and active usage of the platform, M4.4 will develop a framework for monitoring the quality and acceptance of different service categories and repositories, providing a means of actively integrating direct and indirect user feedback into platform development. TA4 will establish a lively exchange with other NFDI consortia (GHGA, NFDI4Biodiversity etc.) to identify common technologies for compute and storage infrastructures and services. Ensuring the long-term sustainability of our NFDI4Microbiota consortium requires us to build and maintain a comprehensive technical platform based on reliable, scalable, and secure software components. Thus, we will apply best practice and state-of-the-art software-engineering techniques such as the five SOLID design principles, rigorous and automated testing, and extensive code reviews. We have implicitly included the additional effort required to achieve this in all our staff calculations.

Fig. 13

5.4.1 M4.1 - Computational infrastructure operations

Contributors: BIBI (lead, 30 PM), EMBL (co-lead, 18 PM), JLU (2 PM), RWTH (sig. in-kind)

Goals: To provide scalable cloud compute resources for the NFDI4Microbiota community.

As a national research data infrastructure, NFDI4Microbiota will offer compute and storage resources to its community members. Currently, the “standard model of computation” in life sciences requires users to first download and install software repositories locally before transferring large amounts of data from public databases. Cloud computing enables economies of scale by pooling compute and storage resources and offering a model where public data is hosted by data providers, allowing users to perform their analyses close to where the data resides. Virtual environments allow maximal flexibility in terms of software stacks, and portable containers enable scientists to share environments or workflows with collaborators and guarantee reproducible research.

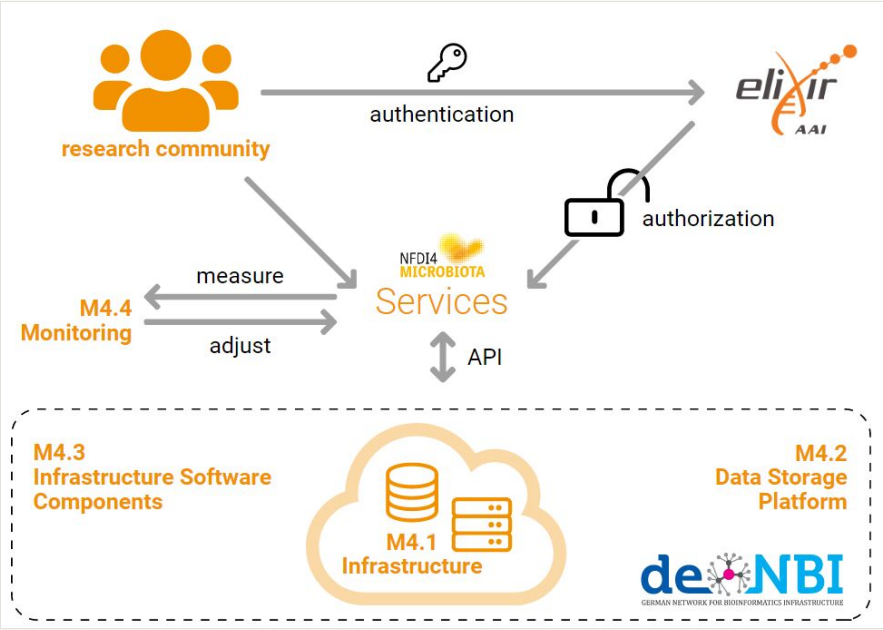

Figure 13. doi

Overview of measures in Task Area 4.

In collaboration with the de.NBI Cloud, we will deploy services and data in scalable cloud computing environments. The de.NBI Cloud, part of the German Network for Bioinformatics Infrastructure (de.NBI), is a German academic federated cloud computing infrastructure for data analysis in the life sciences. As such it provides compute (>30,000 cores) and storage (>40 PB) resources including reference data and relevant bioinformatics methods. Coordination and governance of the de.NBI Cloud take place at Bielefeld University. Together with JLU and EMBL, three of the seven de.NBI Cloud sites are actively participating in the NFDI4Microbiota consortium. As part of ELIXIR-DE, we are already well integrated in ELIXIR and EOSC-Life (Demonstrator D.3, Metagenomics) and will also collaborate closely with other NFDIs such as GHGA and NFDI4Biodiversity.

M4.1 - Computational infrastructure operations will establish, coordinate and operate the technical infrastructure and offer the capacity and expertise to host the computational

infrastructure for NFDI4Microbiota. This includes delivering cloud infrastructure, allocating resources, optimizing performance and capacity, ensuring proper resource management, maintaining availability in order to satisfy the needs and expectations of the community, and meeting service level agreements. de.NBI operates a cloud governance board that supervises the fair, reasonable, and economical usage of cloud resources. Together with M4.4 - Service monitoring & reporting, we will adopt and extend a similar governance structure according to the needs of NFDI4Microbiota. Access to resources will be ensured by a uniform authentication and authorization infrastructure (AAI), developed and maintained by ELIXIR, which is already running very successfully in the de.NBI network. Collaboration is foreseen with other NFDIs from different domains (in this case PUNCH4NFDI) to explore the benefits of establishing an NFDI-wide AAI. Depending on community needs, resources will be provided on different service levels, such as Infrastructure as a Service (IaaS), Platform as a Service (PaaS) and, in later stages, even serverless computing services. Table 30

Table 30.
Milestones to be achieved in measure M4.1 - Computational infrastructure operations.

Milestone	Month	Description
MS4.1.1	6	Initial definition of resource requirements for NFDI4Microbiota services
MS4.1.2	12	Provisioning of prototype infrastructure for NFDI4Microbiota services
MS4.1.3	18	Provisioning of prototype infrastructure for NFDI4Microbiota services
MS4.1.4	30	Provisioning of infrastructure for NFDI4Micobiota services
MS4.1.5	42	SOPs developed for performance and capacity optimization
MS4.1.6	54	High availability protocol established and implemented

5.4.2 M4.2 - Data storage platform

Contributors: JLU (lead, 86 PM), BIBI (co-lead, 20 PM), FSU Jena (6 PM), DSMZ (sig. in-kind), UFZ (sig. in-kind), RWTH (sig. in-kind)

Goals: To develop and provide a scalable and easily accessible domain-agnostic data storage platform.

The core data storage system for NFDI4Microbiota will provide a domain-agnostic, extensible, and scalable platform to store the data of the various user communities and allow fast, efficient and secure access. Essentially, the object storage provided by the de.NBI cloud will be used to store the actual data objects (files). As the requirements and technologies used for this measure are highly similar to the infrastructure to be developed within the NFDI4Biodiversity consortium, and due to several overlaps among co-applicants, an agreement has been made to develop the data storage platform jointly. We have therefore planned additional features and several advanced extensions that will be

implemented primarily within NFDI4Microbiota. By complementing NFDI4Biodiversity's software development plans and combining our forces, we will be able to create a versatile and powerful distributed data storage platform with a comprehensive and tailored set of features, including high availability, optimized georedundant storage of datasets, local and distributed execution of analysis workflows, and automated versioning. Data items will be stored inside the object storage provided by the de.NBI cloud, and we will use a document storage database such as MongoDB for the accompanying metadata. A group of objects will compose a dataset, and a subset of the objects of a dataset can be tagged as a dataset version. Objects can also be aggregated to object groups that act as directories. Objects and dataset versions will be tagged according to semantic versioning principles (e.g. v2.5.27-3.alpha).

The platform will be accessible via standardized and well-documented REST APIs that are provided by automatically scalable Kubernetes deployments. To make optimal use of the scalability provided by the object storage back-end, the actual upload and download of objects, although negotiated via the REST APIs, will be carried out directly via the S3 interface of the object storage. Primarily, the REST APIs will be designed as low-level APIs to be used by software developers who create applications and services for end users such as those described in TA 5. The metadata will be expected in JSON format.

Since the NFDI4Biodiversity consortium will have commenced work on the basic functionality of the storage platform in 2021, it will be possible to provide an early alpha version of the storage platform at the start of the NFDI4Microbiota development phase. Although this early version will be far from feature complete, this should allow for quick and easy adoption of the features that are already implemented, as well as early testing.

The storage platform will consist of two components: (1) a local storage module to handle and manage data storage, and (2) a distributed component to connect locally deployed components. During the implementation of the distributed system, development of the local storage module will continue and new features will be added based on user feedback. We have reached an agreement with the PUNCH4NFDI consortium to continuously share our ideas and experiences relating to cloud storage architectures on a technical level. Initially, we will collaborate on the design of a technical solution to implement tailored caching mechanisms for data and metadata (e.g. based on XCache) and evaluate further possibilities for cooperation. Table 31

Table 31.

Milestones to be achieved in measure M4.2 - Data storage platform.

Milestone	Month	Description
MS4.2.1	4	Local storage alpha deployed for testing incl. CI/CD
MS4.2.2	12	Local storage API stable, start of beta testing
MS4.2.3	18	First stable release of local storage (1.0.0)

Milestone	Month	Description
MS4.2.4	32	Local storage metadata validation
MS4.2.5	24	Distributed storage first alpha deployed incl. CI/CD
MS4.2.6	36	Distributed storage API stable, start of beta testing
MS4.2.7	48	First stable release of distributed storage (1.0.0)
MS4.2.8	52	Local storage data streaming capability
MS4.2.9	48	First local storage ongoing feature extensions and maintenance
MS4.2.10	60	Second local storage ongoing feature extensions and maintenance
MS4.2.11	60	Distributed storage ongoing feature extensions and maintenance

5.4.3 M4.3 - Infrastructure software components

Contributors: JLU (lead, 67 PM), BIBI (co-lead, 15 PM), EMBL (4 PM), UMR (13 PM), UFZ (6 PM), RWTH (sig. in-kind)

Goals: To provide ready-to-use computational frameworks for data analysis and computing services to the NFDI4Microbiota community.

This measure will provide a set of common frameworks to support computational analyses and the development of data processing services such as command-line tools, analysis workflows, and webbased bioinformatics software applications. All frameworks are either required by other measures as IT infrastructure components or are utilized directly as tools by users of the NFDI4Microbiota platform (i.e. usually software developers) to facilitate common computational tasks as well as the implementation and hosting of bioinformatics software tools. We will also provide some general tools for software developers and services to be used for reproducible data analysis, e.g. Juypter notebooks.

Most of these tools will be configurable via our NFDI's online hub to make them easily accessible for non-technical users. The planned tools and services can roughly be divided into three groups: (1) core infrastructure components, (2) database management systems, and (3) tools to support the development of analysis workflows and software applications.

The **core infrastructure components** will be offered as Platform as a service (PaaS) solutions and will be configurable via the NFDI4Microbiota web portal. Many workflow engines still rely on traditional HPC environments for distributing large compute tasks. Parallelization is implicitly defined by the input and output specifications of the individual processes. The resulting workflows are inherently parallel and can be scaled up or scaled out according to the volume of the data to be processed. By using job queuing systems like SGE or SLURM, many computational tasks can be executed in parallel. Based on e.g. the BiBiGrid framework developed at BIBI, we will support simplified and highly automated setups of scalable HPC compute clusters in cloud environments, including an Ansible-based deployment of queuing systems (SLURM) and other analytical frameworks. As a

core part of the technical infrastructure, this measure will also deploy and operate a set of Kubernetes clusters to host the services of this consortium to allow fast and scalable deployments.

The second type of platforms are **database management systems (DBMS)** for various purposes. We aim to support a wide range of database types including traditional relational DBMS such as MySQL and PostgreSQL as well as NoSQL databases such as MongoDB, Apache Cassandra, Neo4j and Redis.

As a third goal, we will make access to **common data analysis tools** easier and help users during the life-cycle of data analysis and software development (M3.2 - Analytical services). Thus, existing software tools that are popular in the user community or those that are requested by participants will be made available in the cloud (e.g. by using container-based deployment), which will facilitate their usage for analyses described in M3.2 - Analytical services. In addition, we will offer analysis notebooks such as Jupyter and Zeppelin as a service.

Within this measure we will also provide a set of tools and interfaces to support stream-based data processing. This will enable direct connections between our central storage platform and local datageneration devices, such as Oxford Nanopore sequencing instruments that are used in larger sequencing centers, smaller laboratories and even in the field. Based on edge computing technologies, client software can be used for local data preprocessing prior to data upload (e.g. to reduce the volume of data). As the preprocessing is use-case dependent, the required preprocessing tools for different analytical workflows will be developed in M3.2 - Analytical services. The upload can also be initiated automatically on the client side; subsequently, further automated and pre-configured data processing steps can be triggered based on the callback system of our storage platform. Table 32

Table 32.

Milestones to be achieved in measure M4.3 - Infrastructure software components.

Milestone	Month	Description
MS4.3.1	3	Initial deployment and start of beta testing of core infrastructure components
MS4.3.2	9	Initial deployment and start of beta testing of PaaS platform with selected DBMS
MS4.3.3	12	Stable availability of core infrastructure components
MS4.3.4	12	Stable availability of PaaS platform with selected DBMS
MS4.3.5	18	First alpha release of authentication proxy
MS4.3.6	24	Annual updates and improvements to core infrastructure components established
MS4.3.7	24	Annual updates and improvements to DBMS PaaS platform established
MS4.3.8	24	Availability of cloud-based deployment of bioinformatics

Milestone	Month	Description
MS4.3.9	28	Stable release of authentication proxy
MS4.3.10	36	Prototype of visualization platform
MS4.3.11	42	Streaming infrastructure stable
MS4.3.12	48	Stable release of visualization platform
MS4.3.13	60	Continuous extension of visualization platform

5.4.4 M4.4 - Service monitoring & reporting

Contributors: BIBI (lead, 30.5 PM), UFZ (co-lead, 12 PM), all other co-applicants with sign. in-kind

Goals: To identify appropriate indicators to monitor the quality and acceptance of NFDI4Microbiota'sservices and repositories, to establish a data-oriented service index, and to measure the impact of ourjoint efforts.

Service monitoring – *i.e.* the continuous recording, processing, and reporting of a defined set of technical and user-based measurements – is nowadays quasi-standard for any productive operating service. It supports a multitude of administrative, developmental and reporting functions during the lifetime of a given service, ultimately increasing its quality, usability and visibility. Key to the whole NFDI initiative is the idea of maximizing community involvement in order to connect data producers with data and service providers and eventually with data and service users. To this end, NFDI4Microbiota will identify and establish service performance indicators on different abstraction levels such that they (i) monitor service life and usage status, (ii) allow services to be adopted and developed according to community-defined needs, (iii) increase the acceptance of FAIR data policies and leverage the usage of FAIR services and repositories, and (iv) provide a measure of the impact of NFDI4Microbiota on national and international communities.

Monitor life status and usage statistics: An initial measure in this task area will be to establish a Service Control Board (SCB) that brings together expertise from all service levels. Its main purpose will be to make sure that agreements on service-related issues are properly addressed in the task areas. The SCB will be open to all consortia members but will initially consist of the leads of M3.2 - Analytical services, M3.4 - Data quality & provenance services, and M4.1 - Computational infrastructure operations, M4.2 - Data storage platform, plus the lead and co-lead of M4.4 - Service monitoring & reporting. The SCB's first task will be to develop a service guideline that outlines a set of minimal requirements such that each service offered by this consortia meets a minimum standard of service quality. This guideline is further complemented by general usage and quality metrics tailored to different service categories.

Develop and adapt to community-defined needs: In addition to the overarching exchange with the microbiota community addressed by the Board for Community Outreach

(BOC, M1.3 - Community outreach and public relations), it is vital to gather user feedback on a per-service basis in order to identify current shortcomings and future needs. To address this task, we will work closely with the SCB to establish a user feedback form for all analytical, data quality, and data storage services as well as the central NFDI4Microbiota web hub. Feedback will be fully transparent to service developers and providers, allowing them to respond to the community in a timely, direct and interactive manner.

Increase acceptance of FAIR data policies: One of the major challenges for all NFDI consortia is to inspire the scientific community to produce, process, and share data in a FAIR manner. The SCB's first job will be to address this challenge by monitoring data usage statistics as an integral part of the data and metadata repositories (as described in M3.4 - Data quality & provenance services and M3.5 - Data deposition & repositories) and linking this back to the service monitoring framework. By doing so, data producers are rewarded by increased usage statistics as a form of a data credit system and, moreover, actively connected to data usage instances that can lay the foundation of future collaborations.

Measure impact and success of NFDI4Microbiota: The SCB will develop and maintain a platform for all NFDI4Microbiota partners and co-applicants and for project governance. This platform will aggregate and integrate all the statistics of the aforementioned measurements and tasks into a single entity with full transparency, making it possible to quantitatively and qualitatively assess the impact of NFDI and NFDI4Microbiota on the research communities. Table 33

Table 33.
Milestones to be achieved in measure M4.4 - Service monitoring & reporting.

Milestone	Month	Description
MS4.4.1	6	Prototype guideline on minimum standard of service quality
MS4.4.2	12	Prototype metrics catalog for different service categories
MS4.4.3	18	Prototype framework for monitoring quality metrics
MS4.4.4	24	User feedback form setup
MS4.4.5	30	Finalized guideline on minimum standard of service quality
MS4.4.6	32	Finalized metrics catalog for different service categories
MS4.4.7	36	Monitoring of data usage
MS4.4.8	48	Single entity monitoring platform established
MS4.4.9	52	Finalized framework for monitoring quality metrics
MS4.4.10	60	FAIR data service index established

Overall schedule, contribution, risks, and dependencies

Fig. 14

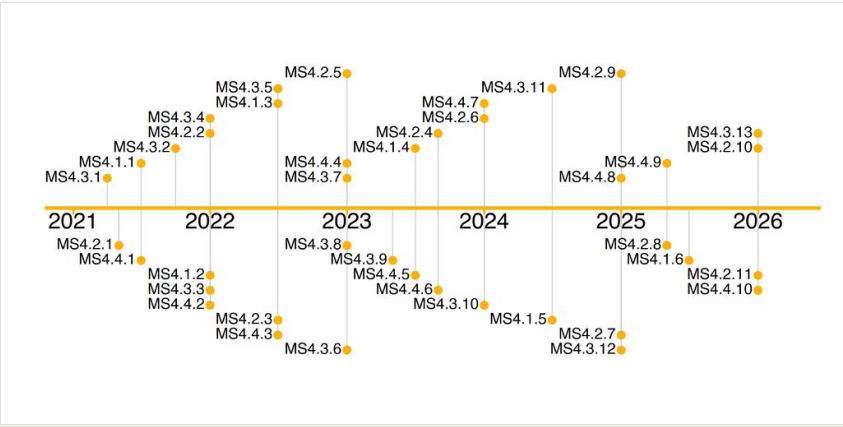

Figure 14. doi

Overall schedule Task Area 4.

Contributions: As TA4 is the backbone of the whole consortium with various closely connected objectives, all contributors within this task area will likewise work in close collaboration and reach out to all the other co-applicants and participants that will actually use the infrastructure. Specifically, BIBI and EMBL will coordinate all efforts regarding cloud infrastructure operations across the different cloud sites, additionally supported by their partner JLU. JLU and BIBI are responsible for the deployment of the data storage platform and will primarily contribute to the storage API and the storage authentication model, supported by DSMZ, UFZ, and FSU Jena. Accompanied by BIBI, JLU will also lead the development of software tools and common components supported by EMBL, UFZ and UMR. This includes generalpurpose platform services and frameworks to facilitate the development, utilization and deployment of analytical frameworks and tools. The definition of appropriate service indicators for monitoring and reporting, their implementation on all service levels, and the development of a monitoring framework will all be coordinated and carried out by BIBI and UFZ.

Dependencies: TA4 will support the integration into the NFDI4Microbiota platform of all software components developed in the measures of this task area and will connect the corresponding services with M5.2 for easier monitoring and reporting.

Risks of implementation and mitigation approaches: (1) Potentially unreliable storage services will be addressed by redundancy of software components and using multiple cloud sites; (2) limited usability based on difficulties in integrating the provided system will be prevented by allocating additional software architect PMs in measures that directly interact with the infrastructure components; (3) disagreement with other NFDI consortia on joint development of platform components will be avoided by defining a minimum set of common platform features and joint adjustments in prioritization of planned developments.

5.5 Task area 5 - Coordination & Communication

Coordinators: ZB MED (lead), HZI (co-lead)

This task area lays the organizational foundations for all the other task areas. Key activities are the definition and establishment of functional (decision-making and advisory) bodies, project governance, and the documentation of processes to provide maximum transparency and ensure efficient operation. The interaction between partners and functional bodies will be facilitated via an internal, GDPR-compliant web platform (extranet) for documents, appointments, and the organization of physical meetings and telephone/video conferences. A smooth workflow will be ensured by regular meetings of the Board of Directors combined with close monitoring of work progress (defined by milestones and deliverables, demand and bottlenecks) and the initiation of any necessary adaptations. The administrative management will be responsible for preparing regular activity reports and summary financial reports including justification of the use of resources and amendments. Furthermore, TA 5 will deliver a sustainable plan for the accomplishment of FAIR principles and open science concepts in microbiota-related systems. Any changes to the work plan or to the financial plan, subject to approval by DFG, will be made by the General Assembly (Organizational structure and viability). Legal issues will be settled in close contact with the NFDI Directorate.

Fig. 15

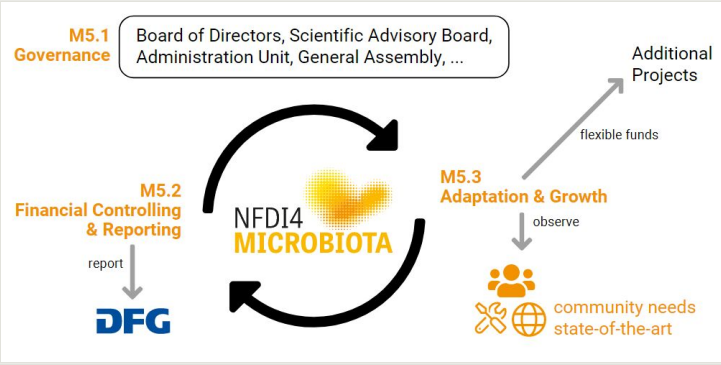

Figure 15. doi

Overview of measures in Task Area 5.

5.5.1 M5.1 - Project governance

Contributors: ZB MED (lead, 24 PM), HZI (co-lead, 24 PM)

Goals: To provide a functional governance structure for this consortium, to establish the consortium's internal bodies, and to bring the voice of the community into the consortium.

The NFDI4Microbiota **Administration Unit (AU)**, supported by ZB MED and HZI administrative staff, will be established to meet these goals. Together with the Board of

Directors (BoD), it will coordinate the establishment and work of the internal consortium bodies. A preliminary Administration Unit, consisting of Alice C. McHardy, Konrad U. Förstner and their project management staff, has already been established to coordinate the writing and submission of this proposal. The Administration Unit will coordinate exchanges between consortium (co-)applicants, as well as other NFDI consortia, the SAB, user council and user community (represented by participants). The Administration Unit will work closely with the BoD to monitor scientific activities within NFDI4Microbiota and release an annual report. This measure also includes drawing up the collaboration agreement between consortium members and suitable by-laws, building on the experience ZB MED gained from drawing up similar documents for NFDI4Health. It will also generate openly accessible documentation of the operations and processes to lay the foundations for clarity and transparency for all stakeholders. To allow for successful and synergistic collaboration between partners across the various TAs, several task groups will be established. Preliminary versions of these have already been formed to work on parts of this proposal. Additional task groups will subsequently be established by the Administration Unit if necessary. Finally, the Administration Unit will explore different options for establishing an NFDI4Microbiota extranet. The most suitable option will be implemented as part of Task area 4 - Technical Infrastructure. This platform will support internal consortium processes and facilitate project management, document and data sharing, and enhanced collaboration between consortium partners. Table 34

Table 34.
Milestones to be achieved in measure M5.1 - Project governance.

Milestone	Month	Description
MS5.1.1	0	Governance structure finalized and Consortium Agreements signed
MS5.1.2	3	Initial meeting plan generated
MS5.1.3	6	First Meetings of BoD and of SAB
MS5.1.4	8	Collaboration agreement finalized and signed
MS5.1.5	13	Process implemented for annual reports on all activities

5.5.2 M5.2 - Project financial controlling and reporting

Contributors: ZB MED (lead, 36 PM), HZI (co-lead, 36 PM)

Goals: To monitor the financial plan for all tasks, organize possible future extensions of the consortium, and report within the consortium as well as to funders.

ZB MED will set up project financial controlling and serve as the central contact point to provide advice on financial matters, supported by HZI. It will prepare management-level justification of resources deployed and the summary financial report. The Board of Directors is responsible for financial reporting to the DFG. The legal basis of financial controlling and reporting will be laid out in a cooperation agreement which will be signed

before the funding period. Annual project costs will be calculated and forecast once a year, in September. If necessary, an application for shifting funds to the following year will be filed. The financial requirements of co-applicants for the coming quarter will be collected by ZB MED on a quarterly basis. ZB MED will request the funds from the DFG and distribute the funds to co-applicants without delay. A financial overview and justification of spending for the previous year will be collected by the ZB MED annually in mid-February and communicated to the funder in March.

For internal reporting, progress within task areas is monitored by the TA leaders and reported by them on a regular basis (semi-annually) to the Board of Directors and the General Assembly. The BoD reports to the SAB. External reporting duties are the responsibility of, and coordinated by, the BoD. The BoD will keep the funder (DFG) and the NFDI Directorate informed of any changes and developments in the project.

Financial plans for including further participants will be drawn up by the BoD and decided upon by the GA. Any necessary revisions of the financial plan will be prepared by the BoD with support from the Administration Unit and decided upon by the GA. In addition, the BoD will organize financial planning and all transactions for internal NFDI4Microbiota project calls.

A mid-term report will be drawn up after three years.

Table 35

Table 35.
Milestones to be achieved in measure M5.2 - Project financial controlling and reporting.

Milestone	Month	Description
MS5.2.1	1	Finance controlling and transfer of funds established
MS5.2.2	9	Regular internal reporting established
MS5.2.3	11	Financial forecasting established
MS5.2.4	18	Mid-term report submitted

5.5.3 M5.3 - Dynamic adaptation and growth

Contributors: HZI (lead, 6 PM), UMR (co-lead, 5 PM), all other co-applicants with in-kinds, ZB MED (285 PM)

Goals: To continuously adapt our strategy and services to meet community needs, offer state-of-theart methodologies and follow best practices, while establishing and maintaining the NFDI4Microbiota infrastructure.

Technology exploration and development of new use cases: The consortium will draw on its close ties with its community to develop new use cases (M1.3 - Community outreach

and public relations). As well as gathering user feedback, we will also take into account the recommendations of the NFDI4Microbiota experts themselves and the NFDI4Microbiota SAB. Existing use cases will be adapted accordingly and new ones developed. Concepts for the required analytical services will be updated accordingly (M3.2 - Analytical services). In addition, the consortium members, being experts in their research areas and well-connected in the community, will periodically assess the state of novel technologies for infrastructure development and emerging data types and decide where the framework can feasibly and sensibly be adapted and extended.

Flexible funding and allocation: To realize the dynamic adaptation of services and support of further data types, develop new training programs, and improve our ability to react to unforeseen technical challenges, we will establish a flexible fund allocation mechanism for a substantial portion (~10%) of the total funding. These so-called Flexible Funds Projects will be awarded annually through a project call process with external peer review. For this purpose, a **Project Review Board (PRB)** – an external body whose members will be selected based on suggestions made by the SAB – will be implemented. It will consist of renowned scientists in changing compositions, based on the topics required for review and excluding members with potential conflicts of interest. The PRB will (1) be convened as needed to review proposals for innovation and implementation projects, and (2) provide an unbiased review of innovation and implementation projects. Based on these reviews, the General Assembly (Organizational structure and viability) will select projects that best serve the overall mission of NFDI4Microbiota.

Growth: We will also explore further incentives for data provision, such as reimbursing data holders for the upload of their data, for instance, depending on the data quality, size, provided metadata, and need for further curation. In addition, we will foster the exchange of data between different platforms by interacting with other NFDI consortia, such as GHGA, NFDI4Biodiversity and NFDI4Health (M1.4 - Connection to other NFDI consortia), as well as international culture collections, highthroughput culturing initiatives (Maier et al. 2018), and microbiome platforms, such as MGnify, MG-RAST, IMG Integrated Microbial Genomes and Metagenomes, and large data-generating consortia such as the Human Microbiome Project (I and II), Tara Oceans and Global Top Soil.

Table 36

Table 36.
Milestones to be achieved in measure M5.3 - Dynamic adaptation and growth.

Milestone	Month	Description
MS5.3.1	5	Concept finalized for Flexible Funds Projects and funding allocation
MS5.3.2	11	First set of new use cases developed
MS5.3.3	25	Annual meeting established to decide on changes to infrastructure and service modifications covered by Flexible Funds Projects

Overall schedule, contribution, risks, and dependencies

Fig. 16

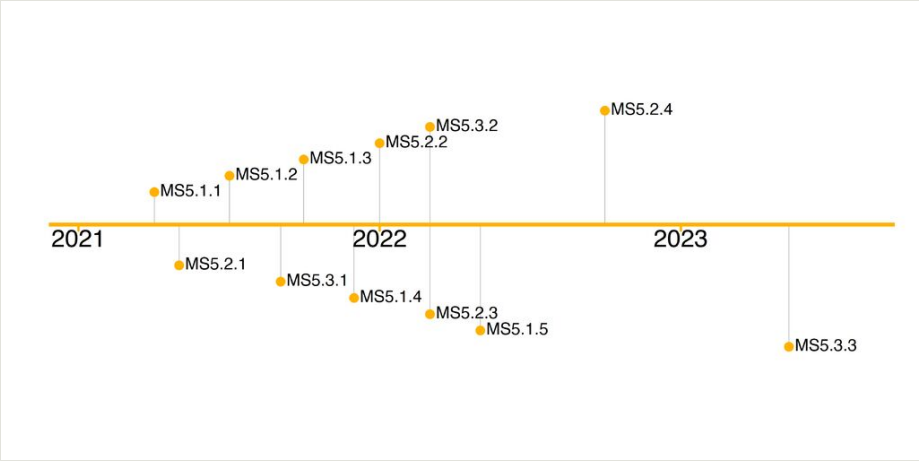

Figure 16. doi

Overall schedule Task Area 5.

Contributions: The aim of TA5 is to provide the organizational and administrative foundations for all other TAs. To this end, ZB MED and HZI – represented by K. Förstner and A. McHardy – will establish the required structures. They will be supported in this task by the Administration Unit. HZI and UMR will jointly work on the adaption of future developments with support from JLU.

Dependencies: The activities of TA5 depend on the cooperation of all other TAs in terms of providing information and reports which are then compiled in overall reports. Additionally, there are numerous external dependencies in the interaction with the DFG and the NFDI directorate.

Risks of implementation and mitigation approaches: (1) Future development might require adaptations of activities and their budgets which could lead to conflicts. These will be addressed by direct communication of the BoD with the affected co-applicant and might be escalated to the GA to make decisions. (2) In the event that parties do not deliver the agreed deliverables in the required quality or time, other parties might take over and funds will be reallocated.

Acknowledgements

For their varied contributions, we thank several people involved but especially (in alphabetical order) James Humphreys, Petra Kneib, and Ulrike Ostrzinski.

Hosting institution

ZB MED - Information Centre for Life Sciences

Gleueler Straße 60

50931 Cologne (Köln)

Germany

Author contributions

All authors contributed to the design of the consortium and participated in writing the manuscript.

Contribution of participants:

- R. Amann: expertise; UC MULTI.
- S. Auer & I. Sens: connect NFDI4Microbiota with NFDI4Ing and NFDI4Chem; identify biotechnologically relevant data; contribute to ontology development and ORKG; harmonize terminology services between NFDI.
- D. Becher: UC MULTI.
- S. Becker: expertise; UC RNA viruses.
- D. Benndorf: support NFDI4Microbiota; initiate collaborations; provide data; UCs MIC-HOST, METABENCH, DataSci, GUT, SYNSYS.
- U.T. Bornscheuer: produce data; contribute to training; UC BIOCAT.
- M. Bott: expertise; UCs BIOCAT, SYNSYS, CRISPR, MULTI.
- A.A. Brakhage: provide insight and virtual infection models; UCs CLINIMIC, GUT.
- T. Chakraborty: UCs PATHO, CLINIMIC.
- T. Dandekar: expertise; initiate collaborations; UC MULTI.
- A. Dilthey: expertise; establish connections with DeCOI; UCs RNA viruses, DataSci, GUT.
- L. Dölken: expertise.
- M. Engstler: expertise; initiate collaborations; UC MULTI.
- L. Falgenhauer: expertise; initiate international collaborations; UCs PATHO, CLINIMIC.
- S. Forslund: participate in R&D efforts and in knowledge and data exchange; make software available; initiate collaborations; UCs GUT, DataSci, CLINIMIC, PATHO, MULTI.
- J. Frunzke: expertise; initiate collaborations; UC CRISPR.
- R. Garrido-Oter: produce and analyze data; high-throughput isolation of microbes; generate collections of sequenced bacterial strains; develop tools; set metadata standards; UC PLANT.
- G. Geisslinger: advice and support; provide network infrastructure; initiate collaborations; UCs RapSARS, RNA viruses, MIC-HOST.

- J. Gescher: provide MinION-Seq data; develop methods and workflows; expertise; UC MULTI.
- H.-P. Grossart: expertise; initiate collaborations; UCs SYNSYS, MICIMG.
- D. Haller: UCs GUT, MIC-HOST, MetaBench, CLINIMIC, SmallPRO, ATTRACTOR.
- S. Hoffmann: expertise on ELN and experience in training; UCs MULTI, DataSci, Soft-Dev.
- M. Höppner: promote NFDI4Microbiota; provide feedback; UC GUT.
- R. Hüttl: expertise; support development and training; UCs SPACE, PHYLOGEN, Soft-Dev.
- A. Kaasch: produce (meta)data; UC RapSARS.
- T. Klassert: produce data; initiate collaborations; expertise; UCs RapSARS, AdaSPat.
- G. Klug: make datasets available; UC MULTI.
- H. Koeppl: expertise; initiate collaborations; UCs SYNSYS, DataSci.
- W. Kühlbrandt: produce and analyze data; provide structures of proteins.
- I. Lagkouvardos: experience; share data; help forming guidelines; share and further develop microbiome data integration solutions; UCs ATTRACTOR, GUT, PLANT, MIC-HOST, MetaBench.
- H. Liesegang: UCs PLANT, DataSci.
- A. Marchfelder: expertise; initiate collaborations; UC CRISPR.
- J. Mattner: expertise; initiate international collaborations; UCs CLINIMIC, GUT, PATHO, MICHOST, MULTI.
- F. Meyer: experience and expertise; initiate collaborations; UCs DataSci, GUT, PLANT.
- R. Möller: UC SPACE.
- F. Narberhaus: analyze RNA-seq data; predict and probe RNA structures; UC MULTI.
- T. Niedermeyer: expertise; initiate discussions; evaluate data; UCs DataSci, MULTI, PLANT.
- G. Panagiotou: expertise; initiate international collaborations; UCs CLINIMIC, GUT.
- K. Papenfort: expertise; UC MULTI.
- M. Schloter: produce highly standardized metagenomics data and sequencing capacity; make a large culture collection available; collaborate on analytical services; UCs SYNSYS, Plant.
- R. Schmitz-Streit: expertise; UCs GUT, MULTI, SMALLPRO.
- J. Schultze: establish easy solutions for submitting data to the NFDI4Microbiota infrastructure; support metadata collection.
- C. Sharma: expertise and deep sequencing technologies; UCs CRISPR, MULTI.
- N. Siegel: expertise; UCs PARA, MULTI, DataSci.
- J. Soppa: expertise; establish connections with the Archaea community; UC MULTI.
- J. Stitz: produce data; UC BIOCAT.
- W. Streit: expertise; initiate collaborations; UCs METAMAR, PLANT, MIC-HOST.
- M.J.G.T. Vehreschild: produce, manage, analyze and integrate data; share (meta)data; help building standards; UCs GUT, CLINIMIC.
- J. Vogel: UCs MULTI, DataSci.

- U. Völker: produce, share and interpret data; UC MULTI.
- L.H. Wieler: UC RapSARS.
- J. Ziebuhr: produce data; detect and analyze viral sequences; UCs RapSARS, RNA viruses.

Conflicts of interest

The authors have declared that no competing interests exist.
Disclaimer: This article is (co-)authored by any of the Editors-in-Chief, Managing Editors or their deputies in this journal.

References

- Amstutz P, Crusoe M, Tijanić N, Chapman B, Chilton J, Heuer M, Kartashov A, Leehr D, Ménager H, Nedeljkovich M, Scales M, Soiland-Reyes S, Stojanovic L (2016) Common Workflow Language, v1.0. figshare https://doi.org/10.6084/m9.figshare.3115156.v2
- Belmann P, Dröge J, Bremges A, McHardy AC, Sczyrba A, Barton MD (2015) Bioboxes: standardised containers for interchangeable bioinformatics software. GigaScience 4: 47. https://doi.org/10.1186/s13742-015-0087-0
- Bierwirth M, Glöckner FO, Grimm C, Schimmler S, Boehm F, Busse C, Degkwitz A, Koepler O, Neuroth H (2020) Leipzig-Berlin-Erklärung zu NFDI-Querschnittsthemen der Infrastrukturentwicklung. Zenodo https://doi.org/10.5281/zenodo.3895209
- Blom J, Kreis J, Spänig S, Juhre T, Bertelli C, Ernst C, Goesmann A (2016) EDGAR 2.0: an enhanced software platform for comparative gene content analyses. Nucleic Acids Research 44 (W1): 22-8. https://doi.org/10.1093/nar/gkw255
- Corrêa FB, Saraiva JP, Stadler PF, da Rocha UN (2020) TerrestrialMetagenomeDB: a public repository of curated and standardized metadata for terrestrial metagenomes. Nucleic Acids Research 48 (D1): D626-D632. https://doi.org/10.1093/nar/gkz994
- Field D, Garrity G, Gray T, et al. (2008) The minimum information about a genome sequence (MIGS) specification. Nature Biotechnology 26 (5): 541-547. https://doi.org/10.1038/nbt1360
- Gonzalez A, Navas-Molina JA, Kosciolek T, McDonald D, Vázquez-Baeza Y, Ackermann G, DeReus J, Janssen S, Swafford AD, Orchanian SB, Sanders JG, Shorenstein J, Holste H, Petrus S, Robbins-Pianka A, Brislawn CJ, Wang M, Rideout JR, Bolyen E, Dillon M, Caporaso JG, Dorrestein PC, Knight R (2018) Qiita: rapid, web-enabled microbiome meta-analysis. Nature Methods 15 (10): 796-798. https://doi.org/10.1038/s41592-018-0141-9
- Haug K, Cochrane K, Nainala VC, Williams M, Chang J, Jayaseelan KV, O'Donovan C (2020) MetaboLights: a resource evolving in response to the needs of its scientific community. Nucleic Acids Research 48 (D1): D440-D444. https://doi.org/10.1093/nar/gkz1019
- Ison J, Kalaš M, Jonassen I, Bolser D, Uludag M, McWilliam H, Malone J, Lopez R, Pettifer S, Rice P (2013) EDAM: an ontology of bioinformatics operations, types of data and identifiers, topics and formats. Bioinformatics 29 (10): 1325-1332. https://doi.org/10.1093/bioinformatics/btt113

- Jaenicke S, Albaum S, Blumenkamp P, Linke B, Stoye J, Goesmann A (2018) Flexible metagenome analysis using the MGX framework. Microbiome 6 (1). https://doi.org/10.1186/s40168-018-0460-1
- Khan FZ, Soiland-Reyes S, Sinnott RO, Lonie A, Goble C, Crusoe MR (2019) Sharing interoperable workflow provenance: A review of best practices and their practical application in CWLProv. GigaScience 8 (11). https://doi.org/10.1093/gigascience/giz095
- Maier L, Pruteanu M, Kuhn M, et al. (2018) Extensive impact of non-antibiotic drugs on human gut bacteria. Nature 555: 623-628. https://doi.org/10.1038/nature25979
- Meyer F, Goesmann A, McHardy AC, Bartels D, Bekel T, Clausen J, Kalinowski J, Linke B, Rupp O, Giegerich R, Pühler A (2003) GenDB-an open source genome annotation system for prokaryote genomes. Nucleic Acids Research 31 (8): 2187-2195. https://doi.org/10.1093/nar/gkg312
- Meyer F, Lesker TR, Koslicki D, Fritz A, Gurevich A, Darling AE, Sczyrba A, Bremges A, McHardy AC (2021) Tutorial: assessing metagenomics software with the CAMI benchmarking toolkit. Nature Protocols. Review Article. https://doi.org/10.1038/s41596-020-00480-3
- Neuroth H, Engelhardt C (2018) Aktives Forschungsdatenmanagement-das DFG-Projekt Research DataManagement Organiser (RDMO). URL: https://opus4.kobv.de/opus4-bib-info/frontdoor/index/index/docId/3688
- RDA COVID-19 Working Group (2020) RDA COVID-19 Recommendations and Guidelines on Data Sharing. Research Data Alliance. https://doi.org/10.15497/rda00052
- Schwengers O, Hoek A, Fritzenwanker M, Falgenhauer L, Hain T, Chakraborty T, Goesmann A (2020) ASA^{3}P: An automatic and scalable pipeline for the assembly, annotation and higher level analysis of closelyrelated bacterial isolates. PLoS Computational Biology 16 (3). https://doi.org/10.1371/journal.pcbi.1007134
- Sczyrba A, Hofmann P, Belmann P, et al. (2017) Critical Assessment of Metagenome Interpretation—a benchmark of metagenomics software. Nature Methods 14 (11): 1063-1071. https://doi.org/10.1038/nmeth.4458
- Wilkinson M, Dumontier M, Aalbersberg I, et al. (2016) The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data 3 (1). https://doi.org/10.1038/sdata.2016.18
- Yilmaz P, Kottmann R, Field D, et al. (2011) Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. Nature Biotechnology 29: 415-420. https://doi.org/10.1038/nbt.1823

Endnotes

*1 https://www.mikrobiologie.uni-jena.de/
*2 https://www.jsmc-phd.de/
*3 https://www.microverse-cluster.de/
*4 https://bic.uni-jena.de/
*5 https://evbc.uni-jena.de/
*6 https://mscj.uni-jena.de/
*7 https://viroinf.eu/
*8 https://digleben.uni-jena.de/
*9 https://www.nfdi4chem.de/

*10 http://www.aquadiva.uni-jena.de/
*11 https://www.microbiomecenter.eu/
*12 https://www.microbiotavault.org/
*13 https://www.sfb1371.tum.de/
*14 https://www.crc1382.org/
*15 https://www.go-fair.org/implementation-networks/overview/fair-microbiome/
*16 https://www.go-fair.org/implementation-networks/overview/discovery/
*17 https://www.ncbi.nlm.nih.gov/sra
*18 https://blog.google/products/search/introducing-knowledge-graph-things-not/
*19 https://www.ebi.ac.uk/ena
*20 https://www.ebi.ac.uk/pride/
*21 https://www.ebi.ac.uk/bioimage-archive/
*22 https://elixir-europe.org/platforms/data/elixir-deposition-databases
*23 http://www.insdc.org/
*24 https://workflowhub.eu/
*25 https://www.researchobject.org/
*26 https://nfdi4microbiota.de/
*27 https://carpentries.org/
*28 https://opensciencemooc.eu/
*29 https://www.wikidata.org/
*30 https://www.orkg.org/
*31 http://www.rfii.de/en/the-council/
*32 https://d-nb.info/1189159759/34
*33 https://www.imngs.org/
*34 https://www.cbd.int/abs/
*35 https://bio.tools/
*36 https://www.livivo.de/
*37 https://nbn-resolving.org/urn:nbn:de:101:1-2020041412321918717265
*38 https://www.publisso.de/en/
*39 https://www.publisso.de/en/research-data-management/rd-documenting/
*40 https://www.bv-brc.org/