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

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

2. Formal details

Planned name of the consortium

NFDI4BioDiversity: Biodiversity, Ecology & Environmental Data

Acronym of the planned consortium

NFDI4BioDiversity

Applicant institution

MARUM – Center for Marine Environmental Sciences, University Bremen, Leobener Str. 8, 28359 Bremen

Prof. Dr. Michael Schulz

Spokespersons

Prof. Dr. Frank Oliver Glöckner, frank.oliver.gloeckner@awi.de, Alfred Wegener Institute - Helmholtz Center for Polar- and Marine Research Bremerhaven

Dr. Michael Diepenbroek, mdiepenbroek@marum.de, MARUM – Center for Marine Environmental Sciences, University Bremen

3. Objectives, work programme and research environment

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

2 Life Sciences (21, 23)

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

Biodiversity is more than just the diversity of living species. It includes genetic and phenotypic diversity of organisms, functional diversity, interactions and the diversity of populations and whole ecosystems. The recent IPBES report has confirmed that we are currently experiencing a dramatic loss of biodiversity. A general understanding of the status, trends, and drivers of biodiversity on earth is urgently needed. NFDI4BioDiversity focuses on the following objectives and task areas:

Promoting research data management as an integral part of biodiversity research: In an ideal world research data management (RDM) is seamlessly embedded within the complete scientific workflow. The insufficient context between basic research and ICT developments has contributed to the fact that large parts of the current data management landscape is fragmented, ephemeral, not efficient, and lacks quality. NFDI4BioDiversity will take care that RDM becomes an integral, funded part of biodiversity research which is a prerequisite for successful data-driven science approaches and a major requirement for the NFDI in general.

Enabling FAIRness of data: As professional data management is still almost absent or underdeveloped in science, it is often done according to individual, not necessarily FAIR compliant workflows. The high diversity in biodiversity and ecological communities aggravates the situation; also, with respect to data sources and formats. A considerable amount of data is collected by citizen scientists and (semi-)professionals which may not even have their data digitized. NFDI4BioDiversity will eliminate deficiencies, inconsistencies, and incompatibilities in data structures and semantics which impede effective re-use of data.

Embedding NFDI4BioDiversity into the national & international landscape of data infrastructure services and science: As science is not limited to national boundaries integration of any future services into the existing international service landscape is crucial. Current efforts are mostly community specific. Generic initiatives like the European Open Science Cloud (EOSC) are urgently needed. They are conceptually well positioned, but still at an early stage of development and not specific for certain disciplines. NFDI4BioDiversity will take care to make best use of existing infrastructures and services and to find a good balance between national developments and the outside world.

NFDI4BioDiversity's activities as part of the NFDI will focus on five task areas:

Task area 1: Prosumer and community engagement (2Involve) - We will engage data users and producers (prosumers) by a range of measures build upon but extending the services supplied by the German Federation for Biological Data e.V. (GFBio e.V.). This includes prosumer and

community tailored training, education and road shows as well as a professional helpdesk and support team. We will take care of any data management requests whereby making use of our network of data managers to expand our front office/back office model. As a large part of all species data (80-90%) is collected by experts and volunteers in natural history societies, NGOs, museums, and citizen science projects, we have measures for a strong co-production approach in biodiversity data management, and also link closely to national and federal state conservation agencies and the environment protection agency (UBA). This task area will include several use cases with community-driven topics and cross-domain user groups (e.g. data science pilots, attractive visualisation, open source analysis tools, data aggregation and linkages to other data).

Task area 2: National and international networking (2Connect) - To increase efficiency, capacities and impact of the NFDI we will work towards ONE NFDI by negotiating concepts and common or complementary measures. This ensures that all cross-domain and generic activities and developments of the NFDI4BioDiversity task areas are coordinated and in line with the overall development of the NFDI. Due to our excellent network (GBIF, GEOBON, CETAF, ELIXIR, de.NBI, other NFDIs etc.) the consortium is ideally equipped to create cooperative infrastructures. Taken together this measure will i) coordinate (inter)national implementations, ii) identify development priorities, and iii) harmonise access to existing services by applying standards at all relevant levels (e.g. data, APIs, registries, certification).

Task area 3: Long term data preservation & publication, certification of tools & services (2Consolidate) - NFDI4BioDiversity will enhance and further develop existing services. This concerns a number of data repositories and service networks (e.g. PANGAEA, GFBio, de.NBI). It meets emerging research items and opportunities, as well as new data, infrastructure, and service requirements in the biodiversity domain. A survey and evaluation of the current data management and data archiving landscape will be followed by evaluation and consolidation activities. A particular focus will be on early data mobilization and the certification of services.

Task area 4: Data integration, exploration, & exploitation - the Research Data Commons (RDC) (4All & 4Future) - We will develop cloud-based infrastructure components and service environments for the integration, exploration, and exploitation of biodiversity relevant data. A major topic will be the “Research Data Commons”, conceived as a virtual expandable infrastructure, collaboratively developed with other NFDI consortia and aligned to the developments of the EOSC. This includes the development of common (meta)data models and standards, cross-domain data integration, governance and support, as well as laying the foundation for the development of a cloud-based “ecosystem” of user supplied applications.

Task areas 5: Collaborative Governance & Sustainability (Coordination) - Building on GFBio e.V., a DFG-funded not-for-profit association set up in 2016, NFDI4BioDiversity will establish a collaborative governance engaging all relevant stakeholders - research institutions & researchers, authorities, citizen scientists, societies, as well as the NFDI directorate, infrastructure & service

providers. Coordinated and responsive structures and procedures will foster a successful and sustainable development and operation of all NFDI4BioDiversity services. NFDI4BioDiversity's budget will include a significant share of money for the flexible support of activities essential for the further development.

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

NFDI4BioDiversity builds on the GFBio project, which addressed key challenges in research data management for biological and environmental data over the last six years. As a federated infrastructure, GFBio comprises large and well-established data centers for nucleotide (EBI) and environmental data (PANGAEA), as well as seven natural science collections for bio- and geodiversity data. PANGAEA is one of the world's largest archives for environmental data certified by WMO, ICSU-WDS, and CTS). The data centers at the natural science collections include the largest German natural history research museums, the network of botanical gardens and the world's most diverse microbiological resource collection, which together not only host more than 75% of all museum objects (150 millions) in Germany and >80% of all described microbial species (Bac*Dive*), but also represent the biggest and internationally-relevant data repositories. A recently joined member of the consortium is the IPK Genebank, Gatersleben, which holds one of the largest collections of crop plants and their wild relatives.

The collaboration with the German Network for Bioinformatics Infrastructure (de.NBI) provides large scale data analysis and cloud storage capacities. Further on, we make use of existing infrastructure providers like the Gesellschaft für wissenschaftliche Datenverarbeitung Göttingen (GWDG), that links our activities to national high-performance computing and large-scale data analysis infrastructures, e.g. the HLRN federation. Similarly, via NFDI4Earth, we expect additional resources from the Helmholtz Federated IT Services (HIFIS). As these partners are infrastructure providers for several consortia, we can thus exploit technical synergies.

During the last years, GFBio partners have developed advanced support services which will be important building blocks for NFDI4BioDiversity: the GFBio Terminology Service (TS) providing a single access point to a variety of biodiversity relevant ontologies and the Visualisation and Analysis Tool (VAT) which allows to visualise, analyse and integrate biodiversity data within an online geographical information system. Among the key services developed and operated by GFBio is a consolidated data submission and brokerage service, which offers a single point of data hand-off to one or several of its data centers. Furthermore, GFBio offers expert support in the preparation of Data Management Plans for DFG grant applications. All services make use of an industry-standard help desk system (Jira) for centralized communication with the users and collect instant user-feedback for each service request.

Further infrastructures and services which NFDI4BioDiversity will made use of are: DataCite, ORCID, Re3Data, CoreTrustSeal, EOSC registries, GBIF, CETAF, GEO-BON, ELIXIR, terminologies (WoRMS, ITIS, ChEBI and more).

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

NFDI4BioDiversity has collaboration agreements with several consortia whereby concentrating on those (1) where we identified essential contributions to the overall NFDI or consortia (2) which complement our portfolio with respect to science and infrastructure.

National Research Data Infrastructure for Earth System Science. NFDI4Earth is a natural partner in terms of common data and metadata models, the re-use of infrastructure services, i.e. in data mobilization and cloud systems. It is a consumer of domain specific e.g. (taxonomic) terminology services and will be a partner for the shared data repository PANGAEA. Data exchange between NFDI4Earth and NFDI4BioDiversity is obvious to e.g. better understand the impact of climate change on biodiversity loss and vice versa.

National Research Data Infrastructure for Agricultural Sciences. With NFDI4Agri there are considerable overlaps in investigated objects e.g. biodiversity of plants or animals. Information about these organisms is stored in repositories, which will be made accessible and usable for mutual exchange. Collaboration is specifically planned in the areas of object identification, phenotypic trait collection and genotyping data management. In NFDI4Agri the data consumers are farmers or breeders which represent highly complementary user communities.

National Research Data Infrastructure for Chemistry. Out of the data managed by NFDI4Chem metabolomics data is of particular interest for the biodiversity community. Integrated data access across the NFDIs is thus crucial. In addition, collaboration is planned on the evaluation and development of data management tools for the early stages of the data lifecycle.

National Research Data Infrastructure for Personal Health Data. Common interests with NFDI4Health are biobanking and sample management as well as common data types like molecular “Omics” data. The legal framework and FAIR metrics that will be developed in the health and medical field in compliance with privacy regulations and ethics principles are of interest for the subset of sensitive data provided by NFDI4BioDiversity e.g. on endangered species. Similar to NFDI4Earth the deeper understanding of wellbeing, resilience and health of each of us is closely linked to the stability of the ecosystem, where biodiversity is a major factor.

Collaboration with KonsortSWD is already projected to exchange regional economic data to contextualise biodiversity studies as well as knowledge on anonymisation. Together with NFDI4Microbiota we are planning to work on common metadata standards for sampling.

We are in contact with NFDI4Crime, BRIDGE4NFDI, DataPLANT, NFDI4BIMP, NFDI4Medicine and NFDI4Life Umbrella to further explore commonalities which will be detailed in the application.

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

- User involvement and adaptive development of NFDI (cross-domain use cases)
- Training, undergraduate and graduate education, professional development
- Reputation and credit systems
- Terminology management and services
- (Meta)data harmonisation, provenance and interoperability across domains
- Standardization (cooperation with standards bodies for technical standards and norms)
- Quality management and assurance - incl. certification of services
- Research data commons - incl. cloud infrastructures, computing power, and AAI
- Governance & sustainability
- Legal questions and frameworks (IPR, Licencing), commercial use of data
- Policy advice and consultation
- International visibility and networking
- Long-term foresight and common strategic planning

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

- **Terminology management and services**

  Terminology services allow semantically enriched data management from retrieval to archiving by integrating and harmonizing heterogeneous terminological resources (incl. taxonomies). In cooperation with all NFDIs we want to extend our terminology service developed for the biological and environmental domain to provide services and tools to find, explore, share and reuse terminologies for the semantic enhancement and harmonization of data across domains. A major point of action, that can only be solved cooperatively, will be the mapping between terminologies from different domains.

- **(Meta)data harmonisation and interoperability across domains**

  Easy exchange of data across NFDIs/domains would leverage new research potential e.g. by combining biodiversity, environmental, and even social science and health data. We will address structural, semantic, and conceptional hurdles for the harmonization of data and metadata. The overall goal is to minimize the necessary efforts. For this purpose, we will concentrate on standards champions in the various domains and to work towards common standards. E.g. following the schema.org principles building on a common core schema with community extensions. For semantic (meta)data harmonization we will need terminology services as described above.

- **Research data commons**

  As part of NFDI4BioDiversity we are planning the implementation of a Research Data Commons (RDC) to pool data and applications for leveraging data science for the different stakeholders. This RDC is conceived as a virtual expandable infrastructure that allows users to store, analyse, share data and results and to combine diverse data types. Together with other NFDIs we will explore if RDC can be extended and serve as a common platform for data sharing and cross-domain data analysis. (Meta)data harmonization as well as terminology management as described above are building bricks for the RDC. Essential parts of the RDC incl. the cloud-based storage layer will be elaborated in collaboration with NFDI4Earth. A particular problem to be addressed when aggregating data from different sources is provenance.

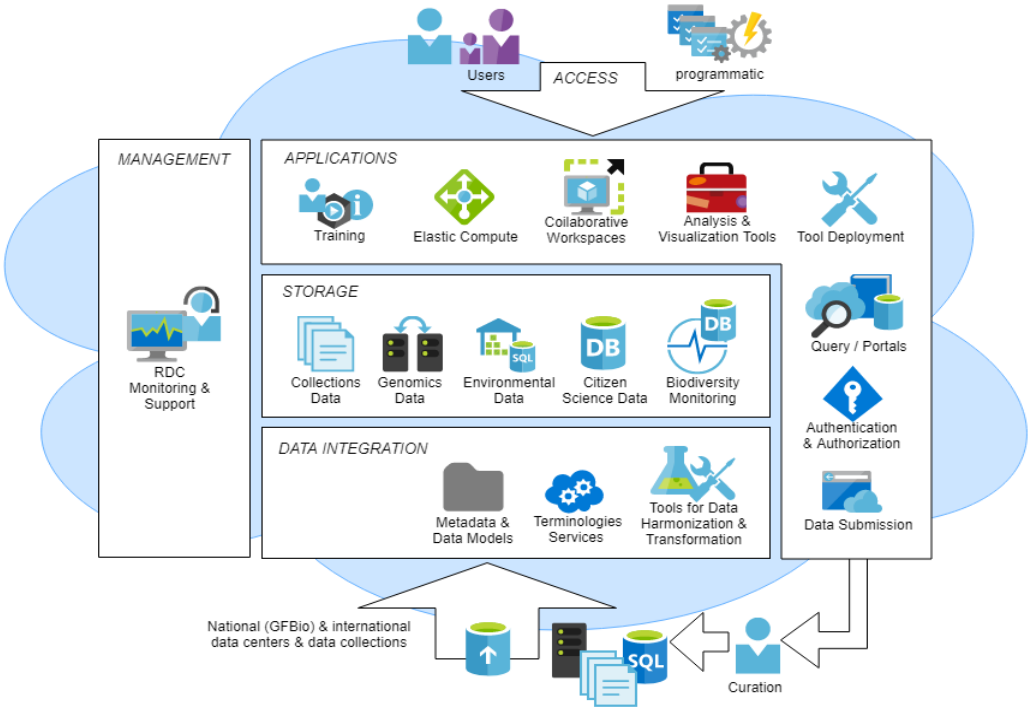

Fig. 1: A Research Data Commons as a common vision for NFDI

- **Certification systems and strategies**

  To meet the increasing demands of funders, publishers, and research organisations to get the quality of data and services formally accredited, certifications by organisations like the ICSU-WDS, DINI e.V., nestor, or the new CoreTrustSeal have become widespread means. Based on our long-term experiences with the certification of the PANGAEA information system, our essential role in the development of the CoreTrustSeal, and our involvement in the H2020 FAIRsFAIR project, NFDI4BioDiversity is in a prime position to share this knowledge, develop strategies for certification and guide the certification process in other NFDI consortia as well. Certification will address authenticity, integrity,

confidentiality, and availability of data and services as well as the assessment of the FAIRness of data centers and their holdings.

- **Graduate School**

  Together with the NFDI4Earth and NFDI4Health consortia and the city and state of Bremen, we are planning to establish a graduate school for research data management and data science.

- **Governance & sustainability**

  Suitable governance structures are a key prerequisite to ensure sustainable operations of a distributed infrastructure like NFDI. Therefore, a major challenge for NFDI and/or the NFDI consortia and directorate will be the identification of an appropriate legal entity which serves the interests of the consortia and service providing host institutions. With the foundation of GFBio association (e.V.) in 2016 NFDI4BioDiversity has ample experience in exploring possible legal forms as well as business models. We would be happy to share this knowledge with all other NFDI consortia to create a common understanding of the pitfalls and challenges of this cross-cutting topic and work towards a common model for governance and sustainability. Together with NFDI4Earth we are currently discussing a ‘collaborative governance’ model.