NFDI4BioDiversity

Key questions/objectives of the consortium

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. Mankind continues to dramatically impact the earth’s ecosystem which is the foundation of human well-being. A general understanding of the status, trends, and drivers of biodiversity on earth is urgently needed to determine management options and devise conservation responses. Answers to scientifically and socially relevant questions can only be found following the FAIR data principles, open science and through availability of data integrated from multiple sources. To foster easy access to interoperable data, NFDI4BioDiversity focuses on the following key objectives:

- Data management as an integral part of research

  Over the last decades on the national as well as the EU level, separate funding streams have been used for Information and Communications Technology (ICT) and Science, resulting in data infrastructure services which are not optimally adapted to scientists needs. Still, within research communities there is a lack of awareness for existing data management services. The situation is worsened by missing incentives in science for the management, archiving, and publication of data.

- FAIRness of data - FAIR+Q

  Besides the fact that data are often scattered across repositories or not accessible at all, the main challenge for integrative studies in biodiversity is the heterogeneity of measurements and observation types, combined with a substantial lack of standard compliance and documentation. Inconsistencies and incompatibilities in data structures, interfaces, and semantics affects data quality and hinders the re-usability. Synthesis as well as hypothesis generation will only proceed if data are FAIR[1] and their quality (Q) can be determined.

- Embedding NFDI4BioDiversity into the national & international landscape of data infrastructure services and science

  NFDI4BioDiversity will cooperate with neighbouring NFDIs like NFDI4Earth, NFDI4Chem, NFDI4Health, NFDI4Agri. NFDI4BioDiversity is a founding member of the NFDI4Life-Umbrella. As science is not limited to national boundaries integration of any future services into the existing service landscape is crucial for success. Nevertheless, current efforts are mostly community specific; generic services are sparse and the impact so far is moderate. Initiatives like the European Open Science Cloud[2] (EOSC) are urgently needed, conceptually well positioned, but are still in an early stage of development and not specified for certain disciplines. NFDI4BioDiversity as part of NFDI as a whole will help to position Germany in the emerging European and international landscape.

[1] The FAIR Data Principles: https://www.force11.org/group/fairgroup/fairprinciples

[2] EOSC: https://www.eosc-portal.eu/

Known needs/current status of research data management in the biodiversity domain:

From a research perspective

Communities dealing with biodiversity and ecological issues are highly diverse with a plethora of taxonomical and ecological specialisations and disciplinary approaches. Standardised data management is often not established yet, and therefore done according to individual, not FAIR compliant workflows. Many data are collected by individual amateurs and (semi-)professionals which may not even have their data digitized. However, assessment of anthropogenic impact on biodiversity change at local, national, or global scale and development of scenarios at different scales needs a sound data foundation with a sufficient coverage in space and time (LTER[3]). Furthermore, biodiversity researchers increasingly integrate various data types ranging from gene sequences, population trends and species functional trait data to estimates of ecosystem structure, functioning and services. This heterogeneity has also been embraced in the development of Essential Biodiversity Variables (EBVs[4]). From the science perspective the compilation and harmonization of distributed data for large scale and/or complex science applications is mostly slow and tedious.

In terms of available information providers and services

There are several key players and information providers in the field of biodiversity and environmental research data. They include national, European and international long-term initiatives (e.g. CoL[5], EoL[6], GBIF[7], GGBN[8], GBOL[9], LTER). They supply useful services and a large amount of reusable data. Nevertheless, they address a rather restricted spectrum of the current biology research data which are structured according the proprietary schemes of their portals.

Over the last five years the DFG funded German Federation for Biological Data[10] (GFBio) project established a national contact point for research data management in biodiversity and environmental sciences. GFBio consists of 20 institutions in Germany, a unique interdisciplinary collaboration between biological and environmental sciences, computer science, and data management communities comprising data centers for molecular (EBI) and environmental data (PANGAEA), as well as seven natural science collections including the largest German natural history research museums, the network of botanical gardens and world’s most diverse microbiological resource collection.

The integration of collection-related, molecular, biodiversity, and environmental data into uniform organisational and technical workflows is a need and a ground-breaking achievement of GFBio. These workflows are based on a set of jointly developed services ranging from standards-based curation, archiving and publication of data to the data portal and visualization functions, as well as terminology services and training units. The GFBio helpdesk supports users with any question concerning research data management and ranging from data management plans, FAIRness of data to legal issues like the Convention on Biological Diversity and the Nagoya protocol.

[3] Long-Term Ecosystem Research: http://www.lter-europe.net/lter-europe

[4] Essential Biodiversity Variables: https://geobon.org/ebvs/what-are-ebvs/

[5] Catalogue of Life: http://www.catalogueoflife.org/

[6] Encyclopaedia of Life: https://eol.org/

[7] Global Biodiversity Information Facility: https://www.gbif.org/

[8] Global Genome Biodiversity Network: http://www.ggbn.org/ggbn_portal/

[9] German Barcode of Life: https://www.bolgermany.de/

[10] GFBio: www.gfbio.org

Summary of the planned research data infrastructure that is specifically intended to address the needs of research users in their respective work processes

Building on the established user community and previous experience of GFBio NFDI4BioDiversity will take advantage of the growing number of NFDI consortia to provide added value, cross-domain services and products. NFDI4BioDiversity will extend its community engagement by including biodiversity monitoring, collection data, as well as systems biology encompassing word-leading tools and collections for FAIR data and quality management. We will provide an attractive and dynamic organizational and technical environment for Citizen Science. The goal to build one comprehensive NFDI includes cross-cutting training and a mutual exchange of data and experiences as well as agreeing on common standards and protocols. To position NFDI4BioDiversity under the NFDI4Life-Umbrella is a logical step in this direction.

The work program of NFDI4BioDiversity is focused along five pillars:

- We will engage users by various activities on different levels like education, training, support, public relations, incentives, reputation and cultural change. We will take care of any requests concerning data management planning, data management, quality assurance, archiving and publication. The consortium will make use of its network of data managers at universities and institutes to expand its front office/back office model.
- We will enhance and further develop existing data management services and tools to meet emerging research opportunities and data requirements in the biodiversity domain. A particular focus will be on early mobilization of data by becoming an integral part of the research process (e.g. by providing data management plans, integrated project data management, software solutions like BEXIS, DWB and Jupyter notebooks).
- We will develop cloud-based infrastructure components and service environments for the integration, exploration, and exploitation of biodiversity relevant data. We will foster community standards, quality management and certification of supplied services. We will proactively engage the user community to build a coordinated data management platform for all types of biodiversity data as a dedicated added value service for all users of NFDI (“research data commons” - aligned to EOSC and other international activities).
- We will address cross-cutting/cross-domain activities in data mobilisation, harmonisation, aggregation and quality management to ensure synergies with related NFDI consortia. Dedicated personnel will be responsible for the mutual exchange of data with NFDI4Life-Umbrella, NFDI4Earth, NFDI4Health, NFDI4Chem, NFDI4Microbiome and NFDI4Agri.
- Finally, we will focus on the governance of NFDI4BioDiversity as part of the overall NFDI and the sustainability of supplied infrastructures and services.

Description of data types

Biodiversity data are highly diverse in terms of data types as well as their spatial and temporal coverages. Typically, data originate from spot measurements to time series (days to decades and data from continuous monitoring) as well from single locations, whole habitats or even large-scale global investigations. They consist of observation data, are results of experiments or modelling, are manually collected or automatically gathered by sensors. Furthermore, a rich set of data and metadata standards exists, fuelled by a high bandwidth of vocabularies and ontologies. Besides this, data formats range from molecular data (e.g. marker genes and ‘Omics) to heterogenous ecophysiological and ecological data, multimedia data (pictures like CT scans, videos, sound files) and taxonomic frameworks up to large scale (terabytes) of global earth observation and streaming data that need to be integrated. An additional source of highly relevant data originates from volunteered science or citizen science projects, which require special care with respect to quality assurance.

In summary, the wealth of data types, formats and sources in the biodiversity domain reflect the range and dynamic of technologies, different levels of biodiversity (from genes to ecosystems) and scientific questions in times of global changes. It is therefore predictable that biodiversity data must be framed e.g. with data from epidemiologic studies or social-economic information, to just name two examples. The integration of heterogeneous data types will be of amble importance for educated decisions of the general public as well as policy makers.

Description of underlying data processing / data analysis methodologies

Today, most data processing on biodiversity data is done individually by researchers. This includes (often manual) data integration across various sources and data cleaning. Often processing is decoupled from data management. Consequently, provenance of integrated data is often not fully available and cleaned data often does not flow back to the original sources. Processing often includes integration with standardized data products like expert range maps, elevation or climate models. Since data is often georeferenced, there is a strong need for powerful visualisations. Nowadays, a large fraction of the data processing is done using R. More and more automatically collected comparatively high-volume data (sensors, remote sensing) requiring stream processing capabilities. Similarly, for molecular biodiversity research bioinformatics pipelines are being used.

Planned implementation of the FAIR principles and information about any existing policies or guidelines in the relevant discipline

In NFDI4BioDiversity the FAIRness of data, in particular the interoperability, is a key requirement for efficient and reliable usage of data from a network of data and service providers. Implementation of the FAIR principles within the consortium is supported by the following instruments and activities:

- Education: (1) summer schools & training, (2) foster data science as part of current curricula (collaboration with NFDI4Life-Umbrella).
- (Early) data mobilization (long tail, citizen science data, data from authorities and learned societies, hidden data from collections, indicator maps[11]). Mobilization will be supported by uniform workflows and tools.
- Incentives and reputation for data producers to support the cultural change (mostly bottom up), in particular pushing the development of data publications as acknowledged scholarly work (e.g. via the RDA Interest Group on Data Publishing[12] and various related working groups, latest: RDA WG data fitness for use[13] and RDA FAIR data maturity model WG[14] and AGU initiative Enabling FAIR Data[15]).
- Certification of the FAIRness of repository services and data holdings, preferably by CoreTrustSeal[16], is already part of the GFBio III work program and the major objective of the H2020 project FAIRsFAIR[17] (NFDI4BioDiversity members participating). Activities are in line with the EOSC perspective (FAIR Data Action Plan[18]).
- Data interoperability will be supported by structural and semantic harmonization of data holdings in NFDI4BioDiversity affiliated repositories. NFDI4BioDiversity will participate in the BiodiFAIRse[19] GoFAIR Implementation Network or operate an own Network dedicated towards the of the biodiversity community. A further activity is the new “RDA

[11] www.ioer-monitor.de

[12] https://goo.gl/AlDbWX

[13] https://www.rd-alliance.org/groups/assessment-data-fitness-use

[14] https://www.rd-alliance.org/groups/fair-data-maturity-model-wg

[15] http://www.copdess.org/enabling-fair-data-project/

[16] https://www.coretrustseal.org/

[17] https://www.fairsfair.eu/ and https://twitter.com/fairsfair_eu?lang=de

[18] https://goo.gl/eLDYTq

[19] https://www.go-fair.org/implementation-networks/overview/biodifairse/

WG Harmonization of measurement and observation types[20]" aiming at an operationalization of terminology services (e.g. the GFBio TS[21]).

Planned measures for user participation and involvement

The NFDI4BioDiversity consortium is representative for the different biodiversity related user communities - universities, research centers (e.g. iDiv[22]), as well as federal and national conservation agencies. 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 will engage in a strong co-production approach in biodiversity data management. NFDI4BioDiversity will thereby foster community building, joint learning, scientific literacy and civic participation in biodiversity data science. Our services will support interoperability with available environmental databases and allow for attractive data visualisations and automated analyses (e.g. de.NBI pipelines - data management by stealth).

With GFBio e.V., NFDI4BioDiversity is in the privileged position to have an active, operational legal entity dedicated to research data management, already in place. As the leading partner of NFDI4BioDiversity, GFBio e.V. will act as the single-contact point for all users, the NFDI4Life-Umbrella, shareholders as well as the NFDI board and panels. GFBio e.V. will take care of any requests concerning biodiversity data management. An important part of the community engagement will also be practiced by a membership in GFBio e.V. With the members meeting being the highest decision body of an association, all members shape directly any decision of NFDI4BioDiversity.

In close cooperation with the NFDI4Life-Umbrella, NFDI4BioDiversity will engage in reputation systems, the cultural change and the injection of research data management as an integral part of good scientific practice in the undergraduate and graduate curricula. Furthermore, we will offer specific training entities and workshops tailored to the needs of our users.

NFDI4BioDiversity will take care about appropriate measures to involve user feedback in the further development of the consortium. This will be tackled by a professional helpdesk, user meetings and instant user surveys to ensure a close connection to the pulse and needs of the users. It will establish an early-warning system that indicates any deviations of the services provided from the user's expectations. User satisfaction will be the guiding principle of NFDI4BioDiversity.

Existing and intended degree of networking of the planned consortium

NFDI4BioDiversity builds on GFBio as a federated infrastructure comprising well established data centers, IT research departments, IT service providers, and biodiversity research institutions. The consortium is engaged in numerous national and international networking activities and collaborations on different levels: policies, education, standards, project data management, infrastructure development. Using this background NFDI4BioDiversity will be firmly embedded into the existing national and international development.

on the national level

NFDI4BioDiversity is in close collaboration with other NFDI consortia: the NFDI4Life-Umbrella is a concerted effort to align, unite and foster the life science communities from the beginning on; NFDI4Earth as a natural partner in terms of common data models, re-use of infrastructure services, i.e. in data mobilization and cloud systems; NFDI4Health with data types (e.g. omics and biobanking/sample management); NFDI4Chem for metabolomics data. Further collaborations (e.g. social sciences) are expected at later stages. The collaboration with the de.NBI network provides large scale data analysis and storage capacities in the cloud. Similarly, we partner with infrastructure providers like GWDG, that link our activities to national high-performance computing and large-scale data analysis infrastructure, e.g. the

[20] https://goo.gl/rTkqbJ

[21] https://terminologies.gfbio.org/

[22] https://www.idiv.de/de.html

HLRN federation[23]. As these partners are likely to be infrastructure providers for several consortia, we can thus exploit technical synergies.

internationally

NFDI4Biodiversity partners are interlinked and well established in relevant initiatives. The list of activities on infrastructures and data services is extensive. E.g. we have stakeholders and partners in EOSC related projects and initiatives, in GBIF[24], DataONE[25], OpenAIRE[26], EUDAT[27], the upcoming eLTER ESFRI[28] as well as we are chairing and contributing to various RDA interest and working groups. We would like to link our own solutions to such offering and influence developments for our community by providing requirements or reference services. Especially, GBIF provides data models and access standards which we already use. Similarly, we have a natural stake in TRY[29], the plant traits database.

between the infrastructure facilities and the research community

NFDI4Biodiversity is well connected to the German and International scientific landscape. This is important to contribute to the necessary policy building to align and incentivize scientists and their projects to adopt our services but also to contribute in the requirement analysis and model building. To this end, we are in close contact with the DFG councils and sections, the learned societies, forums, and associations (e.g. GfÖ[30], GfBS[31], GTÖ[32], VBIO[33], DZG[34], NeFo[35] as well as the Alliance of the Science Organizations in Germany. In addition, we are planning a GFBio III funded workshop to engage scientists and science institutions in the further preparation of NFDI4BioDiversity.

with respect to major networking topic

A major topic will be the “research data commons” to be developed in collaboration with further NFDI consortia. This includes cross-domain data integration as well as laying the foundation for the development of a cloud-based “ecosystem” of user supplied applications. An example for an application is the GFBio VAT system[36], developed with collaboration with GFBio affiliated scientists and meanwhile also used by GEOBON[37].

Name and work address of the contact persons

Prof. Dr. Frank Oliver Glöckner, Alfred Wegener Institute Bremerhaven, Jacobs University Bremen

Dr. Michael Diepenbroek, MARUM – Center for Marine Environmental Sciences, University Bremen

e-Mail: contact@nfdi4biodiversity.org

[23] https://www.hlrn.de

[24] https://www.gbif.org/

[25] https://www.dataone.de/

[26] https://www.openaire.eu/

[27] https://eudat.eu/

[28] http://www.lter-europe.net/elter-esfri

[29] https://www.try-db.org/

[30] https://www.gfoe.org/de/

[31] https://www.gfbs-home.de/

[32] https://goo.gl/L5dPPz

[33] https://www.vbio.de/

[34] https://www.dzg-ev.de/

[35] http://www.biodiversity.de/

[36] https://www.gfbio.org/data/visualizeandanalyze

[37] https://geobon.org/