Letter of intent NFDI4Life Umbrella

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

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

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

2 Formal details

- *Planned name of the consortium*
  NFDI4Life Umbrella research data management infrastructure for life sciences

- *Acronym of the planned consortium*
  NFDI4Life Umbrella

- *Applicant institution*
  ZB MED Information Centre Life Sciences, Gleuelerstr. 60, 50931 Cologne
  Head: Prof. Dr. Dietrich Rebholz-Schuhmann

- *Spokesperson*
  Prof. Dr. Dietrich Rebholz-Schuhmann, ZB MED Information Centre Life Sciences, rebholz-schuhmann@zbmed.de

3 Objectives, work programme and research environment

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

  2 Life Sciences (21, 22, 23)

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

The NFDI4Life Umbrella addresses all overarching needs of research data management in the life sciences with all its domains.

The main objectives of the NFDI4Life Umbrella are:

- Cross-domain interoperability of data sources in the life science domain: Ensuring FAIR (Findable, Accessible, Interoperable and Re-usable) data across life science domain NFDI consortia
- Standardization of cross-domain used data and processes, especially of relevant metadata structures, and harmonization of standards across the life science domain and coordination of the processes, across different NFDI consortia in the life science domain (according to demands)
- Sharing of IT solutions and in particular cloud-based IT solutions developed and provided for the purpose of joint use by life science domain NFDI consortia
- Out-reach to life science domain research communities across consortia through concerted education and training, international representation and visibility of the German life science data community: organ of the life science NFDIs to politics, funders, the Allianz der Wissenschaftsorganisationen, other NFDI consortia, and international organizations such as Go FAIR, RDA and EOSC.
- NFDI4Life Umbrella aims to be a consolidated and strong voice of the German life science communities in national, European and international debates on policies, regulations or standards.
- Solutions for specific life science aspects in generic tasks such as reputation, cultural change or policy making, that can be better addressed in a life science umbrella concept but not addressed in the same way across all NFDI consortia.
- A graduate school will be organised by the members of the consortium to provide training related to innovation driven solutions to PhD students and teach new data analytics methods based on the realm of existing data to the uprising next generation of life science researchers.
- Linking together the existing initiatives with relation to research data management from scientific communities and information infrastructures.

The main task areas are divided in two blocks:

1. Networking and coordination

The NFDI4Life Umbrella consortium is meant to enable cross-cutting support for NFDI consortia that focus on their own communities in terms of research data management, but would require vital support for the integration of their community data into the full range of all communities across the life sciences. It has been already demonstrated that data from different parts of the life science domain (see molecular biology and medicine) is relevant beyond the specific limits of a given part of the life science domain. However each consortium in one particular section of the life science domain encounters overheads in achieving interoperability with the other domains, and all consortia in the life science domain will encounter the same overheads. Reducing these overheads and aligning all the different types of data can be achieved through centralized services (with contributions from the different consortia) and will lead to benefits into the different consortia, e.g. through extended use of the research data across all consortia. Such infrastructures have been envisaged from the ESFRI network ELIXIR, but have only (partially) instantiated in the life science domain (in contrast to other scientific domains).

Specific tasks that will be addressed from the NFDI4Life Umbrella consortium to achieve networking and coordination benefits comprise the following: administering and fostering contacts between the consortia; enabling sustained coordination of joined services, interfaces and standards, e.g. dissemination of terminologies; organization of workshops jointly for members from the life science consortia; active exchange and alignment of guidelines and policies including provision of necessary repository support; collecting the requirements and contributions of the consortia concerning cross cutting topics, sources and solutions;and positioning the NFDI consortia, the semantic resources and infrastructures in the life science community (“outreach”).

2. Infrastructure and services

Data and services are highly intertwined in the life science research domain and the high demand for data analytics required (e.g., for genomics data) initiated the development of cloud infrastructures for public use in the research community. As a result, a distributed cloud infrastructure (see de.NBI https://www.denbi.de) is available in combination with the know-how of data management and data analytics to set the current and future standards for the (re-)use of research data in the research community.

On the other side, this infrastructure setup sets the premises for further growth, further alignment of existing data resources, further reuse of data and tools in a cloud-based infrastructure and forms the central unit for scaling up the overall performance of research communities on a national level.

Specific tasks that would be addressed by the NFDI4Life Umbrella consortium are concerned with the overarching semantic standardization, which comprises the support and supervision of the terminology development (in particular the semi-automatic alignment of terminologies across consortia), the mapping of semantic resources from different domains for better reuse of resources, the automatic adjustment of new semantic resources to existing ones, and the development and provision of the infrastructure for using, access and maintenance of terminologies, ontologies and other semantic resources. Further tasks would be concerned with the indexing, dissemination of and access to metadata, the advanced search for (meta) data (and the ongoing development of advanced searches), and the integration of data with the scientific literature through semantic solutions.

The before-mentioned solutions are geared to serve the scientist to best connect the own data through the existing IT infrastructure with the pool of data and literature in the life sciences domain in an established cloud infrastructure, and by these means to drive the academic prowess of the scientist providing own data, or even better own data in combination with the publication about the data and research, both as active and reusable elements in the cloud-based infrastructure.

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

The German life sciences already feature well advanced infrastructures and networks, like the German National Cohort (NAKO Health Study), the German Network for Plant Phenotyping (DPPN), the life science data management infrastructure FAIRDOM or 'Soil as a sustainable resource for the bioeconomy' (BonaRes). Furthermore the DFG-funded, multidisciplinary consortium German Federation for Biological Data (GFBio,

http://www.gfbio.org) is an existing infrastructure, that has undergone a six year formation process regarding infrastructure and community-building. GFBio follows a holistic approach encompassing technical, organizational, financial, and cultural aspects. To transform the project into a sustainable service infrastructure the charitable association GFBio e.V. has been founded in 2016 as the legal entity. It is now a key service provider for research data management in biodiversity and environmental research acting on the national as well as international level. GFBio might act as a foundational pillar and catalyst for similar processes in other life science subdomains.

Another actor is the platform for Technology, Methods, and Infrastructure for Networked Medical Research (TMF e.V., http://www.tmf-ev.de) which is currently coordinating the Medical Informatics Initiative (MII) together with Medizinischer Fakultätentag (MFT) and Verband der Universitätskliniken Deutschlands (VUD). Its vision is the development and deployment of expert opinions, generic concepts, specimen texts, and IT applications, as well as training and consultation to strengthen the quality and efficiency of medical research and to clarify the legal and ethical foundations for performing medical research. TMF holds long-standing expertise in a range of issues relevant to medicine and healthcare research such as legal or data quality issues.

Additionally, the handling, analysis and storage of enormous amounts of data is a challenging issue across all subdomains in state-of-the-art life science research. Hence, an appropriate IT infrastructure is crucial for performing big data analyses and ensuring secure data access and storage. The cloud infrastructure of the German Network for Bioinformatics Infrastructure (de.NBI Cloud, https://www.denbi.de/cloud) has been established over the last years to enable integrative analyses for the entire life sciences community in Germany and the efficient use of data in research and application.

ZB MED has the national and institutional task to provide access to scientific literature and data in the life science domain. As part of this task, it actively covers tasks such as training in semantic resources (data and metadata management, e.g., for librarians and data scientists), provides an IT infrastructure for access to the scientific literature, to metadata information and to a knowledge environment in the life science domain (increasingly in combination with the cloud infrastructure of the de.NBI site at the University of Bielefeld), and drives research in the efficient use of semantic resources in the domain of text and data mining for the life sciences. Although ZB MED combines different relevant competencies, it heavily relies on research and service partnerships in all domains to support a research infrastructure at scale according to the contributions of ZB MED's partners and ZB MED in itself.

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

Other proposed discipline specific life science consortia agree to join their forces in NFDI4Life Umbrella as coordinative council and declare this structure necessary:

- NFDI4BioDiversity: this consortium anticipates to contribute data about different species and their habitat.
- NFDI4Health: the consortium focuses on data from clinical trial and epidemiological studies and provides such data to the public.
- NFDI4Microbiota: the consortium anticipates to analyse bacterial communities from different environments (including the microbiome), collects and provides such data.

- NFDI Neuroscience: the consortium provides access to data sources from neuroscience research.
- NFDI4AIRR: the consortium standardises the data for immunological responses.
- NFDI4Agri: the consortium collects, standardises and provides heterogenous data from agricultural sciences.
- German Genome-Phenome Archive (GHGA): the consortium enables access to the human genome data from the German genome sequencing centers.

Coordination of the collaborations across the consortia under the umbrella of NFDI4Life enables the different consortia to follow the same data standards and reduces overheads for shared tasks. None of the individual consortia can fulfil these tasks. Furthermore, NFDI4Life Umbrella eases the integrated use of data sources across the different consortia.

Furthermore, the collaboration with other cross-cutting consortia is planned to match the needs of the life science with the special aspects of their topic:

- CompeNDI (Competencies for NFDI) for the topic of education and training
- RSE4NFDI for the topic of research software engineering
- Bridge4NFDI for bridging technological solutions dealing with research data at scale

4 Cross-cutting topics

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

The following cross-cutting topics have been identified from the NFDI4Life Umbrella Consortium (according to a joint workshop in Oct 2018).

Tasks closely concerned with the data from the different consortia in the life sciences comprises the data management planning (and related planning of tools and solutions), namely the Research Data Management Organizer for the life sciences (RDMO4Life https://rdmo.publisso.de/) solution as an approach to work towards a standardized way for the ingestion of high quality research data, but also the use of shared terminological sources, tools and solutions to achieve semantic interoperability across consortia. This work leads to the task for long term preservation of digital data, which includes maintenance, access and reuse of the data. One central piece in this data provision approach form the services for persistent identifiers (PID) that are essential to have the data in the public through the data infrastructure.

As part of the data preservation and data archiving tasks, the legal aspects for data provision have to be kept in mind and addressed through experts in this domain, who would ensure that data is openly available, the licence agreements fit the needs of the researcher and the public alike, and long-term provision is guaranteed from the submission day onwards.

Other specific tasks (from an overarching umbrella NFDI consortium like NFDI4Life Umbrella) would address the research community in the sense that they benefit researchers who want to contribute own data. This contribution includes specific conceptualization, planning, development and roll-out of solutions that would drive the reputation credits of the scientist and would contribute to the culture building for research data management and publication benefits. In essence, delivering own data into a research data management infrastructure should promote the visibility and the scientific credits of the scientist (see PIDs and adjacent

publication for the research data), should enable the scientists to use the open data repository to increase the academic benefits for the researcher, and should serve as a hub for science overall.

All the tasks above have to be accompanied with education and training to achieve a shared understanding of the data infrastructure, where the shared understanding is based on already existing and used semantic resources, tools and solutions, but would also grow with the set of new solutions from the different consortia.

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

Due to the structure of NFDI4Life Umbrella spanning all life sciences, the consortium can contribute as unified voice of the life sciences to all of the above mentioned cross-cutting topics . There will be already an interdisciplinary alignment for these topics that can be rolled out to other disciplines or tested for the whole NFDI.