NFDI4Life Position Paper – Research Data Infrastructure for the Life Sciences

Editorial group: Juliane Fluck, Konrad Förstner, Thomas Gübitz, Birte Lindstädt, Dietrich Rebholz-Schuhmann, Ilja Zeitlin
Contact: contact@nfdi4life.de

1 Introduction

1.1 Background

Data-based research and research data management in the life sciences are increasingly carried out and supported by interdisciplinary research and development teams creating the need for the availability and re-usability of research data from laboratory and environmental studies as well as person-related observations. The FAIR Data principles and the Linked Open Data principles provide a crucial frame for any infrastructure receiving, processing and publishing research data. It is obvious that interdisciplinary teamwork is required to provide and develop research data infrastructures that support the scientists in their daily work and that domain expertise is required to preserve the essence of the data. Furthermore, it is mandatory that any such infrastructure fulfils basic technical needs in terms of usability, archiving, interoperability, accessibility and confidentiality of the data to realise innovation from the reuse of research data.

1.2 Requirements for research data management in the life sciences

In order to operationalise the identification of scientific needs for infrastructural services we distinguish between the more general **domain** life sciences, where biology, veterinary medicine and human medicine, biodiversity, agriculture, environment, epidemiology and nutrition are seen as **subdomains**, and the often interdisciplinarily (i.e. across-domains) working **scientific communities** (e.g. GFBio).[1]

[1] According to a definition of the RfII the "term 'community' refers to a group (association, union) of researchers that are well interconnected socially and follow similar rules of conduct. Communities can form based on a common area of interest (e. g. the 'climate community'), but also based on specific methods (e. g. the 'HPC community'), or even theories (e. g. the 'neurocritical community'). [...] Researchers can belong to numerous communities, and communities themselves can form or dissolve quickly in some cases. [...] In the context of the requirements on information infrastructures, it is not easy to state whether it is better for the measures/services/facilities to address the (smaller, temporary, relatively homogeneous) communities at the level of action or to address the (larger, intrinsically heterogeneous) disciplines and fields. For this reason, the term 'scientific community' frequently used by the German Council of Science and Humanities (or less frequently: 'specialist community') includes both definitions in an intentionally ambiguous manner." (RfII, Position paper - Performance through diversity, p. 74)

DOI: 10.4126/zbmed2018001

The life science research communities bring together a wide range of researchers. Thus different needs in research data management have to be addressed. Also, the status of standardization and processes for publishing and sharing of FAIR data is differently advanced in the various life science communities and platforms. On the other hand, **all life science data can be regarded as interlinked through either underlying fundamentally biological or methodological scientific principles**, i.e. the same biological mechanisms may be relevant in humans as well as in animals or plants, hence in bio-medicine, biological, medical, veterinary, agricultural, environmental, and biodiversity research.
In the life sciences, a large number of data repositories already exists, some of them subdomain-specific such as cohort or population databases in medicine, or crop studies and soil sustainability in agriculture covering certain data types such as gene sequences and next generation sequencing data.

For the Medical Informatics Initiative (MII) for example, standardisation and interoperability of the patients´ medical care data and integration with medical research data and other biomedical data resources such as omics data in data integration centres is relevant and data protection and privacy is mandatory. In the agricultural sciences data is collected to denote the physical appearance of plants and used to judge the state of the development in dependency of environmental and farming conditions. Automatic monitoring of the state of individual plants with their geolocation given is one of the key requirements as well as standard parameters about the climate conditions. Agricultural products serve as the main nutrition source for humans and animals, monitoring the quality of the agricultural products is paramount to increase quality and quantity at the same time. Nutrition data and their influence on health is mainly gathered as observational data. Research in biodiversity monitors species in their natural environments to explore, monitor and eventually understand the relationship between species, their environmental needs and also the environmental benefits. The research communities keep track of species-related features, e.g., phenotypic parameters and the genomics background, in relation to habitat conditions, e.g., by using geolocation as a common anchor. The ongoing research in this subdomain profits, on the one hand, from achievements in agricultural sciences, and on the other hand, contributes to the ongoing research in nutritional research. With respect to the microbiome research on insects, plants, animals and humans it has a profound impact on the wellbeing of any higher organism on earth.

On a broader scale, all life science communities require that existing data resources can be explored for background information, e.g., the repositories on genes and proteins, pathway functions, and the developmental stages of tissues, organs and whole shapes. Ideally, the experimental and observational data can be modelled in a systems biology approach, which is well on its way and requires ever more data for integrative and complex decision making. For example, the modelling of metabolic processes for a given species would bring benefits to the research domains (e.g., agricultural science, biodiversity, nutrition) and altogether would contribute new findings to the medical domain (e.g., systems and precision medicine).
Overall, **there is increasing scientific and societal demand for data and data analysis across communities and subdomains**, for instance between medicine, nutrition and agriculture or between biodiversity, agriculture and anthropogenic influence, for further discovery and understanding of cause and effect chains.

1.3 The NFDI4Life approach

NFDI4Life brings together research communities across the life sciences domain in the context of the planned National Research Data Infrastructure (NFDI), following the recommendations given by the German Council for Scientific Information Infrastructures (RfII).
As a response to the increasing demand for combined data analysis, NFDI4Life brings together scientific communities and research data infrastructures broadly covering the life sciences with particular focus on the subdomains biology, medicine (with veterinary medicine), epidemiology, nutrition, agricultural and environmental science as well as biodiversity research.
Besides providing data resources, as well as data management infrastructure and services for the life sciences in general, with special focus on the subdomains mentioned above, NFDI4Life will also engage in further research based on the produced data (called “data science”) and especially, in new method, standard, quality and process development for research data management and life science data digitalisation.
The integration of data along the life science subdomains requires as a prerequisite making all data **FAIR** (**findable, accessible, interoperable and re-usable**). NFDI4Life is dedicated to lower the barrier of FAIR data generation and re-use for researchers. To reach these goals, NFDI4Life addresses method development, domain-specific and cross-domain standardisation processes and deploys tools to support automated processes and workflows. Together with strategic partners, NFDI4Life will harmonize and further develop the different processes involved in making data FAIR with respect to life science communities as well as to related infrastructures by strengthening partnerships between scientific communities and information infrastructures.
NFDI4Life will enable broad data and information access through advanced information services while addressing data protection issues that are especially important for person-related data, e.g. in clinical research, systems medicine and epidemiology. In these areas, NFDI4Life will support and develop new concepts of accessing this data in protected environments.
Hence, NFDI4Life will foster the further promotion of **good scientific practice in research data management** (transparency, re-usability, reproducibility) along the multidisciplinary field of life sciences and address generic tasks such as building reputation for research data sharing, teaching, training, legal and ethical issues faced across the entire NFDI. It will strengthen communication across communities to overcome the virtual boundaries hampering big data research and innovation. It will act as the driver of an urgently needed cultural change in research data management.
Particular focus in NFDI4Life is placed on **interoperability of data**, the exchange of data and identification of research questions, where transfer of research approaches (as well as improvements) can be achieved based on the interoperability of existing data across all subdomains; such achievements would establish the significant synergies that can only be realised in a consortium covering the broad range of the scientific life sciences. The most prominent example for such cross-cutting approaches is given in the research domain of so called “omics” data. Huge amounts of omics datasets are generated across the life science subdomains which have similar infrastructural requirements in terms of research data processing, deposition, annotation, retrieval and recombination.
NFDI4Life will build on the existing data sources and data management platforms provided by the community-driven data infrastructures and will contribute its own wide range of expertise to explore, shape and exploit the existing data sources and data science approaches for an efficient and meaningful research data management in life sciences.

1.4 Dual strategy of NFDI4Life

The consortium brings together specific expertise from crucial life science subdomains, extending to bio- and medical informatics, but also IT/cloud development, data sciences and data stewardship, semantic technologies (for interoperability), text and data mining, and with similar importance, IT/cloud infrastructures and information sciences as an important component for the archiving and retrieval of the managed data.

The broad coverage of scientific topics across the life science domain provides opportunities and challenges at the same time. The consortium plans to tackle the challenge through a dual strategy**:** it incorporates the advantages of a **broad top down approach** to address **overarching requirements**, i.e., distilling the requirements of the IT/cloud infrastructures and services from the shared demands across the different use cases in the life science domain, and by contrast also through a **bottom up approach** in all subdomains to meet the **subdomain-specific user demands** appropriately.

NFDI4Life will first identify the scientific needs for community-specific as well as generic services in order to provide sustainable state-of-the-art services and to deliver core IT/cloud solutions that can be used across the full range of life science data. In addition, NFDI4Life will build on existing cutting-edge services, adjust them in accordance to the users' needs, develop new services, if needed, and exploit synergies between the members of NFDI4Life as well as other consortia.

As far as NFDI4Life is viewed as a managed process it will support, if required, the constitution of not yet organised communities in the different subdomains of the life sciences in order to help them articulate their infrastructural needs. NFDI4Life will remain open for the involvement of further scientific communities within the life science subdomains and will seek cooperation with other related consortia in the presently arising NFDI structure as well as with other relevant actors such as GO FAIR or the Research Data Alliance.

NFDI4Life will support and drive the development within the whole NFDI process by providing generic modules and services, such as standardisation procedures, strategies for reputation enhancement, training material and training concepts for various target audiences, data protection, and data maintenance. Hence NFDI4Life can make a significant contribution to solving these general issues that all scientific communities are facing.

The particular advantage of a broad consortium is the ability to map and continuously monitor the life science landscape with regard to existing legal, regulatory and funding contradictions and, thus, to assist policy-makers and funders in finding efficient solutions. Since the infrastructural landscape in the life sciences is evolving dynamically, it's important to create an overall coordinating entity in time which would counteract further fragmentation in the life sciences. In this context, NFDI4Life is willing and able to provide policy advice or recommendations about legal and organisational issues.

2 Research data management in life science communities

The high scientific and societal demand for combined analysis of datasets across the life science subdomains is faced with a heterogeneous landscape of related infrastructures. The infrastructures in the different subdomains and related communities vary, e.g. in terms of user involvement, sustainability, level of international integration etc. Moreover, the users or scientific communities themselves may have a different organisational degree or a different practice of research data management.

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). Without claiming to deliver a comprehensive overview of life science communities, we sketch out three further initiatives as examples of life science communities covered by our consortium.

The DFG-funded, multidisciplinary consortium German Federation for Biological Data (GFBio, http://www.gfbio.org) has undergone a five year formation process regarding infrastructure and community-building. GFBio follows a holistic approach encompassing technical, organisational, 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, being a member of NFDI4Life, 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 Universitätsklinika 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 to perform big data analyses 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.

3 Structure and organisation of NFDI4Life

3.1 Involved partners[2] and their research communities

Currently the consortium comprises **university-, Leibniz-, Fraunhofer-, Max-Planck- as well as departmental research** (“Ressortforschung”) **and infrastructure facilities**. Thus NFDI4Life is well integrated within the German science system.
NDFI4Life consists of over 20 research institutions and information infrastructures mainly dedicated to medicine, agricultural and environmental science, nutrition, biodiversity and bioinformatics. The partners comprise information infrastructures focused on a specific subdomain of life sciences and research institutions dedicated to one of the mentioned subdomains. The latter represent the view of their research community while the former are in constant exchange with their user communities. Some information infrastructures and research institutes have a broader focus on several subdomains and/or specialise in method development for research data management.

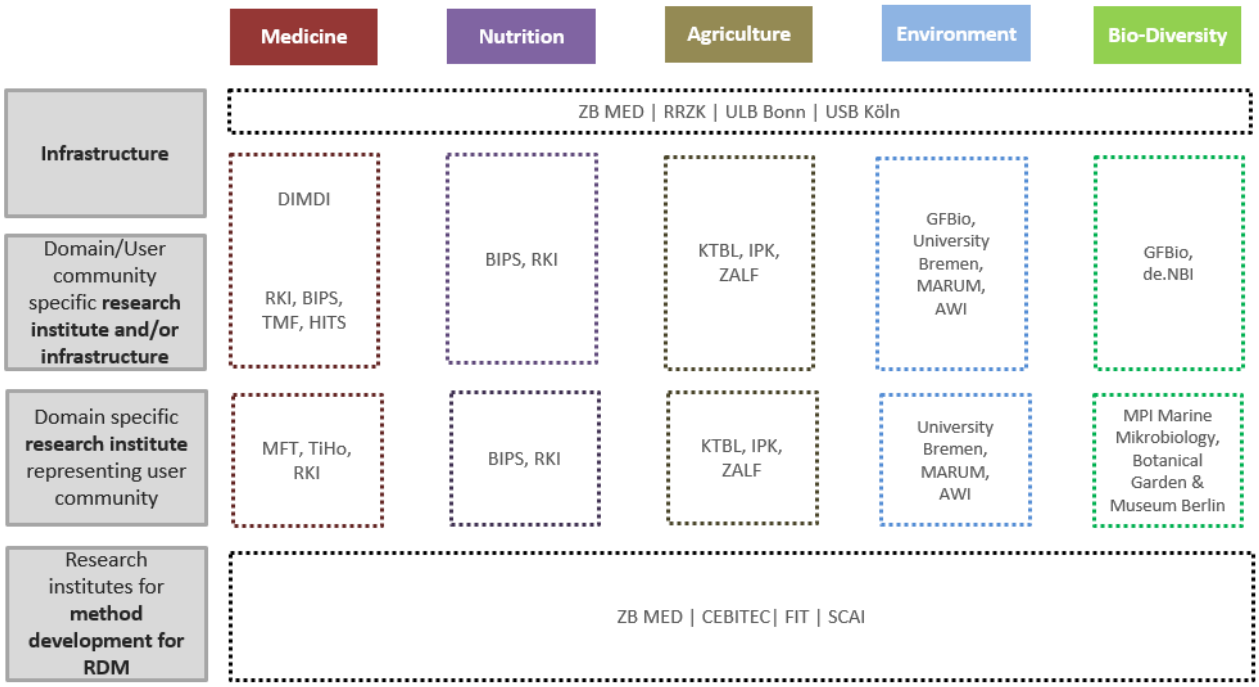

Fig. 1: NFDI4Life consortium partners

3.2 Governance structure and relevant groups

A core bottom-up element in the set-up of NFDI4Life is the representation of data producers and data users. While a number of consortia members are research institutions and data centres and hence represent both sides simultaneously, there will be explicit user representation in the form of an **user advisory board** in order to emphasize the user needs in the building of NFDI4Life. In respect to the broad scientific subdomains in the life sciences it is also necessary to form a **scientific advisory board** (SAB) as a counselling body.

[2] For an overview see appendix.

The internal structure will consist of a coordinating head office and a body as representation of the subdomains medicine, nutrition, agricultural and environmental science as well as biodiversity in NFDI4Life (**steering committee**). As a national infrastructure that overarches all subdomains of life sciences ZB MED - Information Centre for Life Sciences would fulfil the role of the **coordinator** (head office). It has a strong standing as a national infrastructure provider in all areas of the life sciences and a long track record in interdisciplinary and translation work starting from basic biological research to medical, nutritional and agricultural applications. It provides the expertise to integrate and to host literature, numerical as well as semantical data and information. Generic tasks relevant for all subdomains, like legal aspects or reputation system, are already and will be further addressed by **special interest groups** (SIG) within NFDI4Life in coordination and cooperation with other initiatives on national or international level.

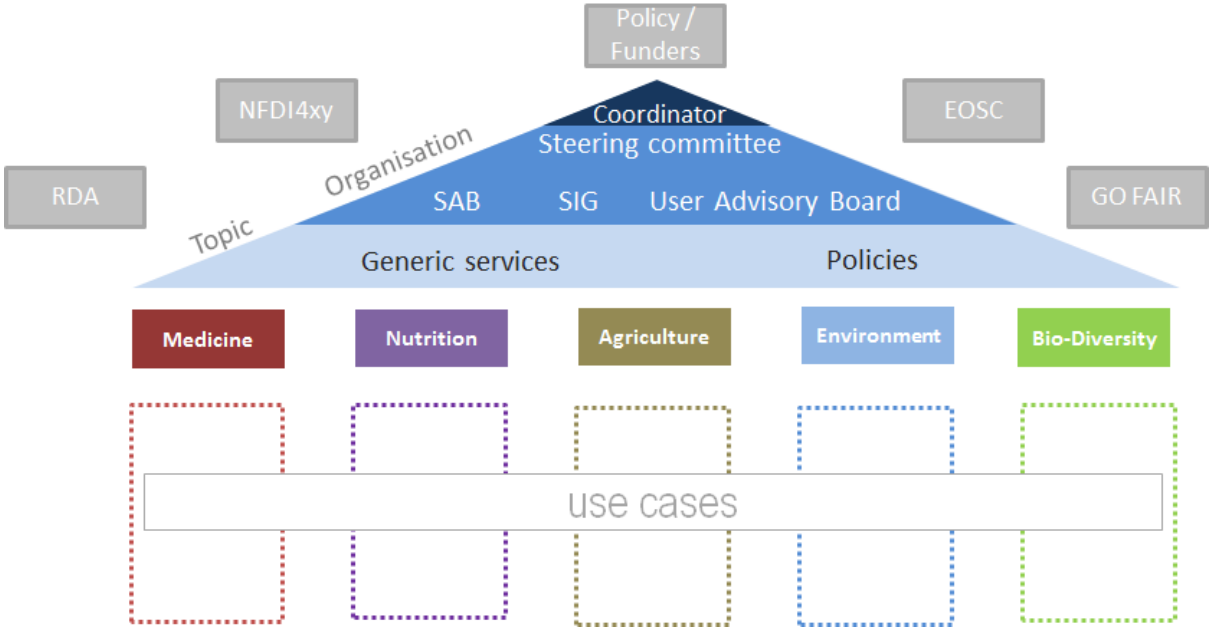

Fig. 2: NFDI4Life coordinating structure and relevant groups/actors/initiatives

3.3 User-driven development of research data infrastructures

Research data management requires not only expertise in data handling but needs close contact between data experts and the researchers generating and using the data in order to understand their needs. Thus, domain expertise as well as community networking and a process for monitoring user needs is necessary to address data producers and users properly.
The consortium partners are well accepted in the different communities and have set up the already existing research data infrastructures within their research communities. They will continue to provide and develop domain-specific services – with respect to parallel developments and needs of the other user communities.
NFDI4Life has already set up **special interest groups** (SIG) with experts from all subdomains to build a close matrix structure in order to address subdomain-specific demands, current developments and to agree on overarching standards. The SIGs cover the fields of action: standards - metadata, ontologies, identifier, etc., information services, teaching and training, methods for FAIR research data ecosystem, reputation through research data, archiving, publication, integration of data and privacy/sensitive data.

Furthermore, to make different user needs and required developments more concrete and to make the impact of NFDI4Life measurable and adjustable, the consortium worked out eight **use cases** which are *interdisciplinary* in character, each involving several *scientific communities*:

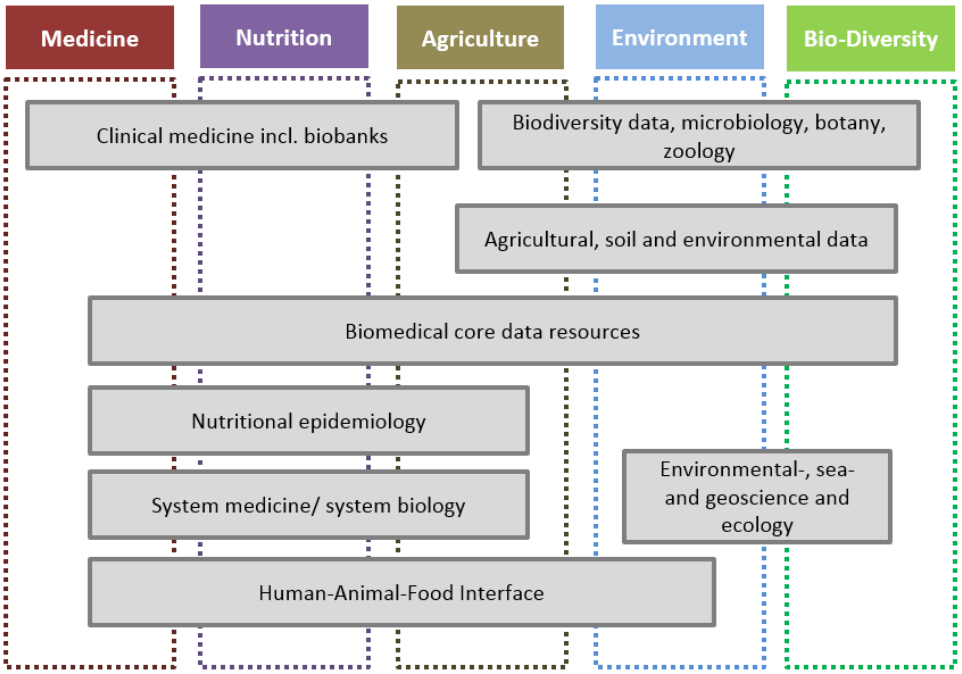

Fig. 3: NFDI4Life Use Cases

The above mentioned use cases will serve as science demonstrators for unleashing the synergistic potential of interdisciplinarity by triggering collaboration between involved actors. Based on these use cases, NFDI4Life will demonstrate the added value for data owners and data users. In doing so, the more advanced partners in terms of research data management are supposed to act as catalysts for others, for instance through the identification of data champions or data ambassadors. The use cases are supposed to build on scientific communities as well as on existing infrastructures.

4 Added value and impact of NFDI4Life

NFDI4Life understands its role as a model for future consortia and the success of NFDI as a whole. NFDI4Life members are ready and strongly committed to contributing actively and to achieving fast results.

NFDI4Life will

- provide science-driven, sustainable and state-of-the-art services in research data management.
- support researchers via provision of standards, advanced and user-driven IT development, professional guidance and training in the data generation and publication process, including the provision of training concepts for different target audiences in the life sciences, like researchers, technicians, information infrastructure staff, students, and emerging fields such as data stewards, data librarians.
- enable a culture of improved data management for data producers and data consumers, and make sure that the FAIR Data principles are implemented.
- empower scientific communities/scientific users to articulate their infrastructural needs and therefore place them in a crucial role in its governance structure.
- serve as an emergent consortium that will enable interdisciplinary cutting-edge research along the multidisciplinary field of life sciences.
- realise synergies for life sciences researchers in terms of (generic) infrastructural services for RDM, hardware availability, specific training in generation and use of research data, legal questions, informed consent, data protection issues and policy processes.
- assure data, software and infrastructure quality by providing and fostering quality processes and certification mechanisms.
- provide quantitative and qualitative measurement instruments of impact to policy-makers and funders.
- be a consolidated and strong voice of the German life science communities in national, European and international debates on policies, regulations or standards.
- open up international resources for German researchers and make German research internationally more visible.

Acknowledgement

We would like to thank all contributors for their input to this position paper and the work done in the consortium: Wolfgang Ahrens, Oya Beyan, Constanze Curdt, Stefan Decker, Michael Diepenbroek, Jens Dierkes, Janine Felden, Frank Oliver Glöckner, Martin Golebiewski, Anton Güntsch, Uwe Heinrich, Dietrich Kaiser, Ulrich Lang, Sabine Leonhard-Marek, Daniel Martini, Ulrich Meyer-Doerpinghaus, Wolfgang Müller, Hubertus Neuhausen, Iris Pigeot, Annette Pollex-Krüger, Uwe Scholz, Sebastian C. Semler, Henriette Senst, Alfred Pühler, Frank Wissing.

Appendix

Partners

1. **BIPS - Leibniz Institute for Prevention Research and Epidemiology**
   https://www.leibniz-bips.de
2. **Botanischer Garten & Botanisches Museum Berlin - Freie Universität Berlin**
   https://www.bgbm.org/en/research
3. **CeBiTec - Center for Biotechnology**
   https://www.cebitec.uni-bielefeld.de/
4. **de.NBI - German Network for Bioinformatics Infrastructure**
   https://www.denbi.de/
5. **DIMDI - Deutsches Institut für Medizinische Dokumentation und Information**
   https://www.dimdi.de/
6. **Fraunhofer FIT - Fraunhofer-Institut für Angewandte Informationstechnik**
   https://www.fit.fraunhofer.de/
7. **Fraunhofer SCAI - Fraunhofer-Institut für Algorithmen und Wissenschaftliches Rechnen**
   https://www.scai.fraunhofer.de/
8. **GFBio - Gesellschaft für Biologische Daten e.V.**
   https://www.gfbio.org/de/gfbio_ev
9. **HITS - Heidelberger Institut für Theoretische Studien**
   https://www.h-its.org/de/
10. **KTBL - Kuratorium für Technik und Bauwesen in der Landwirtschaft e.V.**
    https://www.ktbl.de/
11. **IPK - Leibniz-Institut für Pflanzengenetik und Kulturpflanzenforschung Gatersleben**
    www.ipk-gatersleben.de/
12. **MARUM - Zentrum für Marine Umweltwissenschaften**
    https://www.marum.de/
13. **Max-Planck-Institut für Marine Mikrobiologie**
    https://www.mpi-bremen.de/
14. **MFT - Medizinischer Fakultätentag**
    https://medizinische-fakultaeten.de/
15. **RKI - Robert Koch-Institut**
    https://www.rki.de/
16. **RRZK - Regionales Rechenzentrum der Universität zu Köln**
    https://rrzk.uni-koeln.de/
17. **TiHo - Stiftung Tierärztliche Hochschule Hannover**
    https://www.tiho-hannover.de/
18. **TMF - Technologie- und Methodenplattform für die vernetzte medizinische Forschung e.V.**
    www.tmf-ev.de/
19. **ULB - Universitäts- und Landesbibliothek Bonn**
    https://www.ulb.uni-bonn.de/
20. **USB Köln - Universitäts- und Stadtbibliothek Köln**
    https://www.ub.uni-koeln.de/
21. **ZALF - Leibniz-Zentrum für Agrarlandschaftsforschung e.V.**
    http://www.zalf.de/
22. **ZB MED - Informationszentrum Lebenswissenschaften**
    https://www.zbmed.de/