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

Data in PLANT research

Acronym of the planned consortium

DataPLANT

Applicant institution

Albert Ludwig University of Freiburg (UFR)
D-79104 Freiburg

Head: Prof. Dr. Dr. h.c. Hans-Jochen Schiewer

Spokespersons

Dr. Dirk von Suchodoletz, Computer Center, Albert Ludwig University of Freiburg,
dirk.von.suchodoletz@rz.uni-freiburg.de

3. Objectives, work programme and research environment

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

Biology [Bioinformatics (201), Plant Science (202)]

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

In modern hypothesis-driven science, researchers increasingly rely on effective research data management services and infrastructures that facilitate the acquisition, processing, exchange and archival of research data sets, to enable the linking of interdisciplinary expertise and the combination of different analytical results. The immense additional insight obtained through comparative and integrative analyses provides additional value in the examination of research questions that goes far beyond individual experiments. Specifically, in the research area of **fundamental plant research** that this consortium focuses on, modern approaches need to integrate analyses across different system levels (such as genomics, transcriptomics, proteomics, metabolomics, phenomics). This is necessary to understand system-wide molecular physiological responses as a complex dynamic adjustment of the interplay between genes, proteins and metabolites. As a consequence, a wide range of different technologies as well as experimental and computational methods are employed to pursue state-of-the-art research questions, rendering the research objective a team effort across disciplines. **The overall goal of DataPLANT is to provide the research data management practices, tools, and infrastructure to enable such collaborative research in plant biology.** In this context, common standards, software, and infrastructure can ensure availability, quality, and interoperability of data, metadata, and data-centric workflows and are thus a key success factor and crucial precondition in barrier-free, high-impact collaborative plant biology research. Toward this, the **key objectives** pursued by this consortium are:

1. A specific **community standard** for fundamental plant research (meta)data and workflow annotation, based on generic, existing and emerging standards (e.g., ISA model, MIAPPE) and ontologies in plant science.
2. **Assistive mechanisms and services** to build, link and maintain the complete research context during data acquisition, curation, analysis, and publication.
3. **Mechanisms for collaborative research** based on enrichment and automatized crosslinking of plant-research specific (meta)data to facilitate research context management.
4. A **cloud-based open reference implementation** of these mechanisms and services, and a central hosted instance thereof.
5. A robust, **federated infrastructure** both for data computation and management covering the complete data lifecycle.
6. **Comprehensive training** of community members through workshops and summer schools and providing open training material.

We propose an organization of the necessary work into several closely coupled task areas organized around plant researchers and following a workflow-centric, bottom-up approach:

Task Area 1 (Quality, Interoperability, and Standardization) will work towards developing the envisioned plant-research (meta)data standards. We believe that to ensure that standardization

is effective with respect to ensuring data quality and data/workflow interoperability, an integrative effort is needed that considers these aspects simultaneously through:

- A user- and workflow-centric analysis of metadata and definition of their standards requirements across a wide national and international userbase.
- Ensuring FAIR principles of data, such as development of strategies for data discoverability through automated harvesting of metadata and indexing, providing an enhanced search and access interface (role based, provide search functionality based on domain specific semantics, across data from different repositories, bioinformatic databases, and services), annotation, and standardized research data.
- Evaluation and creation of community-specific measures to enable and foster long-term FAIR adherence of community data via DataPLANT resources. Development of plant research-specific data quality metrics (following the AMPLE principle).

These efforts will be conducted to strengthen and coordinate standardization efforts in plant research-related data and workflow annotation and will be closely linked with other relevant NFDIs nationally, and e.g., ELIXIR, EOSC, EMPHASIS, ePLANT, and MIAPPE internationally.

Task Area 2 (Software, Service, and Infrastructure) is aimed at providing software tools, software services, and infrastructure services for (meta)data, and workflow creation, management, sharing, and evolution providing the basis for collaborative plant research. This will entail:

- Further development and maintenance of long-term, sustainable storage and access to data and associated methods and workflows, as well as reproducibility, including certification of the appropriate environments (especially, built-in adherence to FAIR principles through the use of infrastructure).
- Development and dissemination of novel methods for data intensive computing in computational plant biology.
- Interfacing with vendors to develop and/or improve open, FAIR data formats from the point of measurement.
- Development and establishment of open management services for workflows as software-as-a-service (DataPLANT SaaS) and a central instance of these services.
- Development of financing / business models to ensure sustainability and maintenance of infrastructure, services and software.

These work packages will provide improvements to data and workflow management across the entire lifecycle of plant research (meta)data.

Task Area 3 (Transfer, Application, and Education) will focus on developing mechanisms for interaction and education with stakeholders (plant researchers) and community-building towards furthering collaborative research in plant biology. These efforts will be directed towards:

- Building on the successful de.NBI education services and fully established training courses and channels, developing new training programs for specific user communities in data and workflow standards and management, data literacy, scientific data analysis, and computational methods, in the context of the to-be-developed specifications and infrastructures. This includes both education of young researchers as well as the ongoing qualification of researchers and practitioners in plant biology.
- Building communities through active communication of developed standards, platforms and infrastructure resources.

- Application of the objectives to a bioinformatic research infrastructure.
- Implementation of the developed standards, software, and infrastructures at and beyond participating research centers through partnering in international communities.
- Develop and implement mechanisms for interested parties beyond the participating institutions to participate in DataPLANT in all areas.

As an overarching goal, these measures will grow awareness of project efforts and goals to ensure maximum relevance to a large and international community of plant researchers. **All areas** address cross-cutting aspects and include networking within the NFDI on corresponding topics.

Fig.1: DataPlant is designed to be user centric. All TaskAreas are directed towards the needs of the plant researcher as data champion. Training and application (TaskArea III) ensures the usability in practice and will lead to the formation of a central information resource for fundamental plant research.

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

In DataPLANT we aim to establish a homogeneous interconnected infrastructure environment to enable modern plant research at the highest level, building on an existing base infrastructure (bwHPC (BinAC), bwSFS, de.NBI cloud, bwCloud) and tailoring resources to the fundamental needs of plant researchers. The by-now significant infrastructure for bioinformatics in Freiburg and Tübingen grew out of local cooperation of groups nearly two decades ago, leading to federations at the state level and cooperation at the federal level. Groups of researchers in different settings pooled resources with funds from the state and federal ministries to create an increasingly diverse landscape of machines and services. Both the community and the service providers learned that having an established core of personnel and services as well as operating a relevant infrastructure starts to attract additional users and grant money. With the rising number of users, we started to create flexible governance bodies and established means to match the financial resources brought in to provide storage space and compute services. Most of the investments are still mainly project driven, but the level of coordination in planning and extension of the infrastructure and services is increasing. The development of sound, long-term refinancing models is still an open issue to be tackled in cross-cutting NFDI activities.

Based on these experiences, our approach is to fuse existing capabilities and provide the user with information on which data repository, ISA template, minimum information requirements and (cloud) computational framework best fit the individual research step and data type within the current project, and interconnect the researchers via a central management service. The envisioned service will be able to recommend annotations and heavily reduce repetitive information submission across different data within a project. Big data objects and compute instructions (common workflow language, biocompute objects) are linked, ensuring identity within the project and allowing layering on top of existing data and compute infrastructures that already operate as state / federal resources and are supported by research efforts underway among the consortium participants (e.g., Galaxy and nf-core). These tools and services foster the standardization process and provide the necessary foundation to couple future activities towards (meta)data enrichment and data context continuation.

Infrastructure is at the core of complex scientific workflows and sound data management. We follow a holistic approach, spanning data acquisition, various levels of computation, selection, combination, enrichment, and publication. We view unified permanent storage and computation as a crucial precondition for avoiding gaps in workflows and results. (Meta)data will be preserved and augmented through all relevant stages at every stage of processing. This will build on the established MIAPPE standard which is accepted in the community and is being co-developed by several of the participants.

The fusion of infrastructures across the consortium will address the varying demands of users and lower the barrier of entry to data-collaborative science for new projects and users who are inexperienced in this area (e.g., junior scholars). Thus, the return on investment will scale with increasing interconnectedness of infrastructure, user support, and education; these are among the central goals of the planned consortium. Only an NFDI consortium appears able to provide the fundamental framework to develop the necessary standards, training and support procedures, and infrastructure coupling and synergistic value for data-driven plant research and will open the door to new forms of computational research workflows (e.g., involving machine learning).

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

The DataPLANT NFDI combines corresponding expertise of the initially proposed BioDATEN4NFDI and the DaPLUS consortium (see corresponding extended abstracts submitted to the 1st NFDI Conference), making expertise in data analysis and management available to the fundamental plant research community. In general, DataPLANT will provide a gateway to plant research data, ensuring open standards according to FAIR principles. Inherent to this is a (meta)data standardization process based on international standards and rules, and hence seamless national and international interoperability and interfacing. Due to the consortium's focus on fundamental plant research, data collected from our data champions in the plant community will become a resource for plant research in general.

Essential insights gained from fundamental plant research are ultimately transferred towards applied plant research. Therefore, we plan a close collaboration with NFDI4Agri (agricultural science) at a very early stage to ensure compatible standards and barrier-free exchange.

Also, in the context of (meta)data standards for omics data we envision a close collaboration with NFDI4Microbiota. Additionally, there is shared interest in (meta)data modelling and exchange of ideas and concepts to orchestrate, run, and govern a federated infrastructure with the Text+ consortium.

DataPLANT supports cloud-based infrastructures, in particular the Research Data Commons as conceived by NFDI4BioDiversity. In this regard we also intend to collaborate with NFDI-Neuro to improve generic data workflow management. On the infrastructure level cooperation with other service provides like the RHRK and other research institutions computer centers is envisioned. While focusing on the omics data in plant research, image data resulting from phenome studies are envisioned to be handled in close collaboration with the technology-specialized consortium NFDI4BIMP. They will provide generic and domain-spanning tools and services for the storage and management of microscopy and photonics-based imaging data.

However, success in data management strongly depends on user effort and data literacy, rendering training and education essential. Therefore, a general comprehensiveness of universal techniques on how to handle data has to be conveyed during early education. DataPLANT aims at a wide-ranging training that embraces consortia in different domains in life sciences such as NFDI4Agri, NFDI4BioDiversity, NFDI4Neuro, NFDI4BIMP, and NFDI4Microbiota. These activities will prominently include various forms of e-learning, summer schools and workshops. In addition, we will provide training courses on how rich plant metadata can be used for building hypotheses. Also, we will offer Galaxy training and qualification, allowing the plant community to leverage large computing power for questions they could not run on their own hardware. Joining forces with NFDI4Chem, the exchange of basic FDM and molecule-specific training materials between the initiatives is planned.

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.

Cross-cutting topics of relevance for DataPLANT will be:

- FAIR data principles
- User participation and dynamic development of NFDI and its (meta)data standards
- Reputation and credit systems
- Quality management
- Education and Training
- Interoperability and Transfer
- Research data commons (hard- and software infrastructures)

- Preserving research context
- Data and Metadata Provenance
- Governance & Sustainability
- Careers
- (Cross-domain) Search & Indexing

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

We consider all the following topics to be of a substantially cross-cutting nature; it appears sensible and beneficial to consider these topics beyond the boundaries of the planned consortium and work towards NFDI-spanning solutions, providing a common framework but allowing for specialized solutions where necessary. We believe that many of the below topics should be considered at the consortium level initially but should be worked on at a higher level during the initial phase of the NFDI. We will actively contribute our insights to these discussions.

** FAIR data principles:* Fair data compliance surely is a key objective in all NFDI consortia and will ultimately lead to interoperability across domains. The DataPLANT approach delivers FAIR compliance by design: (F) A core database ensures persistent and unique identifier for all entries. (A) DataPLANT is building a layer on top of trusted infrastructure and repositories that themselves ensure accessibility. (I) We develop a dedicated version control semantics based on current data standards and formats, ontologies and information requirements to ensure interoperability. (R) Due to the proposed tracking and linking strategy, we gather accurate information on provenance. Further, DataPLANT will enable exporting snapshots of the project using Research Object as an international exchange format for guarantied reusability. Automatization of FAIR compliance that minimizes the user effort might be extendable to other NFDI consortia. We would be very interested to discuss and contribute towards NFDI-spanning solution.

** User participation and dynamic development of NFDI and its (meta)data standards:* User involvement and motivation is essential for a successful NFDI. Novel approaches to motivate the user for appropriate data management need to be discussed and evaluated across all consortia. Automated data curation based on different research contexts, will allow a cross benefit between users to encourage user participation. Our domain specific user community is thematically coherent and interconnected via the collaborative research environment of DataPLANT. By linking analysis and compute platforms to current data, we will be able to recommend processing and analysis approaches that suit the (experimental) data. This will add additional value to appropriate annotation for the researcher that recorded the data by facilitating the data processing. Here, we exploit the fact that data annotations in plant research do not need a high level of anonymity, compared to e.g., medical data. Tracking incremental changes and data

source identity will allow the system to automate data curation based on expert knowledge linked across all projects in the plant domain.

* *Reputation and credit systems:* For scientists in all domains it is of utmost importance to get credit for their work. In DataPLANT there is a special focus on storing and preserving the complete research context. Providing infrastructure and tooling covering the complete research cycle from data generation to publication, DataPLANT can ensure (meta)data and software provenance and keeps track of owner(s)/user(s) using ORCID. In this scenario we can generate DOI to enable citations for data resources, especially in data-aggregation scenarios, and include information of all participants during the research process at the particular state of snapshot for publication. Reaching compliance in this regard across the NFDIs will enable science open data and data publication as a relevant part of scholarly communication.

* *Quality management:* Verification improves the long-term reproducibility of research results from digital workflows, since these are not solely dependent on the data and its quality-assured description. We will adapt basic quality checks based on cross-referencing and common rules, but also use MIAPPE meta-data to allow for single data point and data-set outlier detection as the overall data body grows. Additionally, preserving processing software including the entire configuration and specific environment of software libraries, operating system, and underlying hardware is also part of the quality management of data. The envisioned service in DataPLANT can be used to reproduce certain software environments automatically. This will allow for automatized testing (equivalent to unit testing) on the data processing level. Further, it seems to be necessary to develop NFDI-wide standards to access the software quality itself, following guidelines already established in the open source community (e.g., JOSS compliance).

* *Interoperability and Transfer:* By working in and contributing to standardization groups and by relying on widely used container formats such as ISA-Tab, we will broaden interoperability between different NFDIs and internationally. At the same time, we will use the mandate of the fundamental plant science consortium to help improve standards and formats to adapt to the need of the plant researchers.

* *Research data commons (hardware and software infrastructures):* There will be a common ground for all NFDI consortia regarding base level infrastructure and demand for sustainable operation for both storage and compute infrastructure to host higher-level/domain-specific services. In DataPLANT, the infrastructure providers bring in long-term experience in federated infrastructures, expertise in setup of cooperations, and governance for cooperatively organized services. The BinAC high performance computing cluster caters especially to the demands of the bioinformatics community within the bwHPC context. To complement HPC services, the federated bwCloud was introduced and significantly pushed Galaxy services. Close coordination and cooperation with the bioinformatics community convinced the de.NBI consortium to host services in sizable proportion in Freiburg and Tübingen (infrastructure- and software-as-a-service clouds). These services are already open to the German bioinformatics community

enabled by a common AAI. A new addition and complement to the service stack is bwSFS, Storage-for-Science, supporting the Freiburg and Tübingen HPC cluster communities. The Freiburg and Tübingen computer centers host significant installations already open for many domains including other NFDI consortia.

* *Preserving research context:* In pretty much every field preserving just the data objects risks losing significant properties of the research context. Thus, data, workflow software, and the base-level technology stacks need to be considered in a joint context. Considerations for a long-term re-use or the provision of research (software) environments and secured software sources is still in its infancy. The consortium brings in a strong team working on sustainable long-term access for over 15 years. Concepts and practice of software citation have been developed, as well as guidelines and infrastructure to manage and preserve software dependencies which could made available to all NFDI. However, with technical progress and especially the advance of virtualization, container, cloud and related technologies, research environments became interconnected and interactive, and research data and software intertwined, such that access and re-use is only possible if all components are available and usable. In order to adapt FAIR data principles, especially to ensure long-term re-usability of a wide variety of research outputs, novel methods are required for all NFDI, and to be integrated in research data management strategies.

* *Provenance:* For all scientific communities, the origin of data is factor essential to trust. Especially when using data records of third parties, transparent knowledge of the creation processes, applied quality assurance, review processes, also in connection with the organisation of origin, processors and curators, is of utmost importance. This also includes the documentation of the handling of any raw data and the applied pre-processing, until data is available in a format suitable for analysis. The DataPLANT consortium addresses workflow documentation and hierarchical provenance schemas since the analyses are performed in different automatic preprocessing and filter processes. Many findings will be directly transferrable to other NFDI as well. It makes it possible to make any change of a data set transparently, which allows both desired and possibly lossless changes, such as format conversions, as well as collaterals and modifications, such as those caused by incorrect algorithmic processing or data transmission errors.

* *Governance & sustainability:* Large scale cooperation requires the balancing of needs and interests of the partners as well as the moderation of conflicts in all domain specific NFDI and on the general level. Gains and benefits from well-maintained and annotated data enjoyed by one group do not necessarily and directly link to costs spent and efforts provided by another group, refinancing of infrastructures, and support personnel. In DataPLANT we bring in the infrastructure provider's long-term experience in federated infrastructures, and expertise in setting up cooperations and governance for cooperatively organized services. Thus, a key commitment of the consortium is to build effective governance and control structures for the NFDI to reconcile the interests and aspirations of the community and infrastructure/service providers. From the very

beginning, they should actively regulate the business relationships after consolidation. This means developing communication and organisational structures, enabling effective exchange of information between the participating institutions and research groups. In parallel, processes for the comprehensive participation of user groups in corresponding decision-making processes will be developed.