NFDI Proposal

DFG form NFDI 110

Contents

1. General Information

1.1 Name of the consortium in English and German

DataPLANT: *Data in PLANT research*

DataPLANT: *Daten in Pflanzen-Grundlagenforschung*

1.2 Summary of the proposal

In modern hypothesis-driven research, scientists increasingly rely on research data management (RDM) services and infrastructures to facilitate the collection, processing, exchange, and archiving of research records. RDM enables the combination of interdisciplinary expertise, as well as comparison and integration of various analysis 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. The central aim of the DataPLANT project is to advance this added value in the field of basic plant research. Specially, in fundamental plant research, the (molecular) principles of plant life are investigated, which determine plant growth, crop yield and biomass production. The methods used for this purpose, from transcriptomics, proteomics and metabolomics to imaging techniques, produce high-dimensional polymorphic data that must be integrated for meaningful interpretation. Successful collaboration and use of data of different modalities – from many sources and experiments, pre-processed or analysed with a variety of algorithms – requires contextualization of the data. The FAIR Data[1] and Linked Open Data Principles provide critical guidelines for RDM. Various consortia have therefore made proposals for best practice and compliance with these principles, but it is almost always the initiative of individual researchers to implement them. Therefore, comprehensive information on the required quality for use by third parties is rarely

available. Researchers have been shown to require practical assistance in exploiting the fragmented and complex resource landscape. This increases the need for a tailor-made (infra) structure for RDM. By combining technical expertise in the fields of fundamental plant research, information and computer sciences and infrastructure specialists, DataPLANT will support plant scientists in every RDM concerns. DataPLANT will create a service environment to contextualize research data according to the FAIR principles with minimal additional effort and to support the entire research cycle in modern plant biology. The tailor-made service landscape in DataPLANT will consist of technical-digital assistance as well as on-site personnel assistance. DataPLANT thus creates a central entry point and a valuable subject-specific data and knowledge resource. In combination with teaching and training concepts, data literacy is strengthened and a long-term motivation for the creation of well-indicated data objects is generated. By integrating plant science into the NFDI network as a whole, DataPLANT is driving the digital transformation and democratization of research data in the field.

1.3 Zusammenfassung

In der modernen hypothesen-basierten Forschung sind Wissenschaftler zunehmend auf Dienste und Infrastrukturen für Forschungsdatenmanagement (FDM) angewiesen, die die Erfassung, Verarbeitung, den Austausch und die Archivierung von Forschungsdatensätzen erleichtern. Dabei ermöglicht FDM erst die Verknüpfung von interdisziplinärer Expertise, sowie Vergleich und Integration verschiedener Analyseergebnisse mit dem darauf beruhenden immensen zusätzlichen Erkenntnisgewinn. Das Ziel des Projektes DataPLANT ist es, diesen Mehrwert für den Bereich Pflanzen-Grundlagenforschung zu avancieren. In der Pflanzen-Grundlagenforschung werden die (molekularen) Prinzipien des pflanzlichen Lebens erforscht, die Pflanzenwachstum, Ernteertrag und Biomasseproduktion bestimmen. Die hierzu eingesetzten Methoden von Transkriptomik, Proteomik und Metabolomik bis hin zu bildgebenden Verfahren erzeugen hochdimensionale polymorphe Daten, die verarbeitet und zusammengeführt interpretiert werden müssen. Erfolgreiche Zusammenarbeit und Nutzung von Daten unterschiedlicher Modalitäten – aus vielen Quellen und Experimenten, vorverarbeitet oder analysiert mit einer Vielzahl von Algorithmen – erfordert eine Kontextualisierung der Daten. Die FAIR Data and Linked Open Data-Prinzipien bieten entscheidende Richtlinien für FDM. Verschiedene Konsortien haben daher Vorschläge zur besten Vorgehensweise und Erfüllung dieser Grundsätze gemacht, doch ist es fast immer an der Initiative der einzelnen Forscher, diese auch umzusetzen. Daher stehen umfassende Informationen über die erforderliche Qualität für die Verwendung durch Dritte oft nur in selten Fällen zur Verfügung. Es hat sich gezeigt, dass Forscher praktische Unterstützung bei der Nutzung der fragmentierten und komplexen Ressourcenlandschaft benötigen. Dies erhöht die Notwendigkeit einer maßgeschneiderten

(Infra)struktur für FDM. Durch den Zusammenschluss von technisch-fachlicher Expertise in den Bereichen Pflanzen-Grundlagenforschung, Informations- und Computerwissenschaften und Infrastrukturspezialisten wird DataPLANT Pflanzenwissenschaftlern im Umgang mit Forschungsdaten individuell angepasst unterstützen. Dabei wird DataPLANT eine Serviceumgebung schaffen, um Forschungsdaten nach den FAIR-Prinzipien mit minimalem Zusatzaufwand zu kontextualisieren und den gesamten Forschungszyklus in der modernen Pflanzenbiologie zu unterstützen. Die maßgeschneiderte Servicelandschaft in DataPLANT wird sich aus technisch-digitaler Assistenz sowie personelle Vor-Ort-Assistenz zusammensetzten. DataPLANT schafft so einen zentralen Einstiegspunkt und eine wertvolle fachspezifische Daten- und Wissensressource. In Verbindung mit Lehre und Trainingskonzepten wird das Sachverständnis im Umgang mit Daten gestärkt und eine Langzeitmotivation zur Schaffung wohlannotierte Datenobjekte erzeugt. Durch die Integration der Pflanzenwissenschaft in das Gesamtnetzwerk NFDI, treibt DataPLANT den digitalen Wandel und die Demokratisierung der Forschungsdaten im Feld voran.

1.4 Applicant institution

Applicant institution	Location
Albert-Ludwigs University of Freiburg (UFR) Head: Prof. Dr. Dr. h.c. Hans-Jochen Schiewer	Fahnenbergplatz, 79104 Freiburg

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

1.5 Co-applicant institution

Co-applicant institutions	Location
Technical University of Kaiserslautern (TUKL) Head: Prof. Dr. Dr. h.c. Helmut J. Schmidt	Erwin-Schrödinger-Straße 52 67663 Kaiserslautern
Jülich Research Center (FZJ) Head: Prof. Dr.-Ing. Wolfgang Marquardt	Wilhelm-Johnen-Straße 52428 Jülich
Eberhard Karls University Tübingen (EKUT) Head: Prof. Dr. Bernd Engler	Wilhelmstraße 5 72074 Tübingen

Co-spokesperson	Institution, location	Task area(s)
Prof. Dr. Björn Usadel	IBG-4 Bioinformatics, Jülich Research Center	Task Area I
Dr. Jens Krüger	High Performance and Cloud Computing Group, IT Center, Eberhard Karls University Tübingen	Task Area II
Jun. Prof. Dr. Timo Mühlhaus	Computational Systems Biology, Technical University of Kaiserslautern	Task Area III

1.6 Participants

Participants	Institution (where applicable), location
Prof. Dr. Rolf Backofen	Bioinformatics, Albert-Ludwigs University of Freiburg
Dr. Olaf Brandt	Head of IT at University library, Eberhard-Karls University of Tübingen
Prof. Dr. Andrea Bräutigam	Computational Biology, Bielefeld University
Prof. Dr. Stefan Deßloch	Heterogeneous Information Systems, Technical University of Kaiserslautern
Dr. Marianne Dörr	University Librarian, Eberhard-Karls University of Tübingen
Prof. Dr. Alisdair Fernie	Central Metabolism, Max-Planck-Institute of Molecular Plant Physiology
Prof. Dr. Christoph Garth	Scientific Visualization, Technical University of Kaiserslautern
Dr. Björn Grüning	Bioinformatics, Albert-Ludwigs University of Freiburg
Dr. Petra Hätscher	University Librarian, University of Konstanz
Prof. Dr. Eric Kemen	ZMBP, Eberhard-Karls University of Tübingen

Prof. Dr. Dr. hc. Edda Klipp	Theoretical Biophysics, Humboldt University of Berlin
Prof. Dr. Ute Kraemer	Molecular Genetics and Physiology of Plants, Ruhr University Bochum
Dr. Daniel Lang	Plant Genome and Systems Biology, Helmholtz-Zentrum München
Prof. Dr. Dario Leister	Plant Molecular Biology/Botany, Ludwig-Maximilians University of Munich
Prof. Dr. Heike Leitte	Visual Information Analytics, Technical University of Kaiserslautern
Prof. Dr. Klaus F.X. Mayer	Plant Genome and Systems Biology, Helmholtz-Zentrum München
Dr. Sven Nahnsen	QBIC, Eberhard-Karls University of Tübingen
Dr. Anja Oberländer	Head of Open Science, Communication, Information, Media Centre (KIM), University of Konstanz
Dr. Klaus Rechert	Longterm access, Albert-Ludwigs University of Freiburg
Prof. Dr. Ralf Reski	Plant Biotechnology, Albert-Ludwigs University of Freiburg
Dr. Inga Scheler	Regionales Hochschulrechenzentrum Kaiserslautern, Technical University of Kaiserslautern
Prof. Dr. Karl Schmid	Crop Plant Biodiversity and Breeding Informatics, University of Konstanz
Jun. Prof. Dr. Sandra Schmöckel	Physiology of Yield Stability, University of Hohenheim
Prof. Dr. Gerhard Schneider	Prorector, Albert-Ludwigs University of Freiburg
Prof. Dr. Waltraud Schulze	Plant Systems Biology, University of Hohenheim

Prof. Dr. Thomas Walter	Computer Center, Eberhard-Karls University of Tübingen
Prof. Dr. Andreas P.M. Weber	Institute of Plant Biochemistry, Heinrich Heine University Düsseldorf

1.7 Subject orientation of the proposed consortium

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

2 Consortium

2.1 Research domains or research methods addressed by the consortium, objectives

Plants use the energy from the sun to produce living matter driving all our ecosystems. They are primary producers in natural and agricultural settings and support most other life forms, either directly or indirectly. Thus, society depends on plants as sources of our energy, nutritious food, of sustainable materials and fuels, and of medicinal compounds. It is consequently vital that we understand the fundamental processes that determine plant growth, crop yield and the production of biomass. Plants are sessile and not able to escape their surroundings. Hence, they are often challenged by stresses such as disease or climate change limiting their growth and decrease crop productivity. Within the frame of their genetic capacity, plants are able to perceive and respond to changes in environmental conditions. These responses represent a complex dynamic interplay between genes, proteins and metabolites and are manifested processes on all systems level. **Fundamental plant research** is the study of these fundamental processes to improve our understanding of the molecular basis of plant life. The goal is to elucidate the underlying physical and chemical principles of how a plant functions on a mechanistic molecular level. Therefore, fundamental plant research is a multidisciplinary research domain, including **molecular genetics, biochemistry, cell biology, systems biology, physiology, development and evolution**, and that finds application of important discoveries in **plant biotechnology** and **plant breeding** (DFG 202-[01, 04, 05, 06, 07]). With a clear domain specific emphasis, the central approaches in fundamental plant research to dissect the underlying principles and elucidate the functioning of plants by: (i) recording multiple parameters under changing conditions, (ii) measuring the effect of genetical and biochemical manipulation to alter gene or protein activity, (iii) and analysing natural genetic diversity and evolution. In 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. Phenotyping platforms and high-throughput technologies such as **mass spectrometry, next generation sequencing, spectroscopy, and imaging techniques** are used to simultaneously

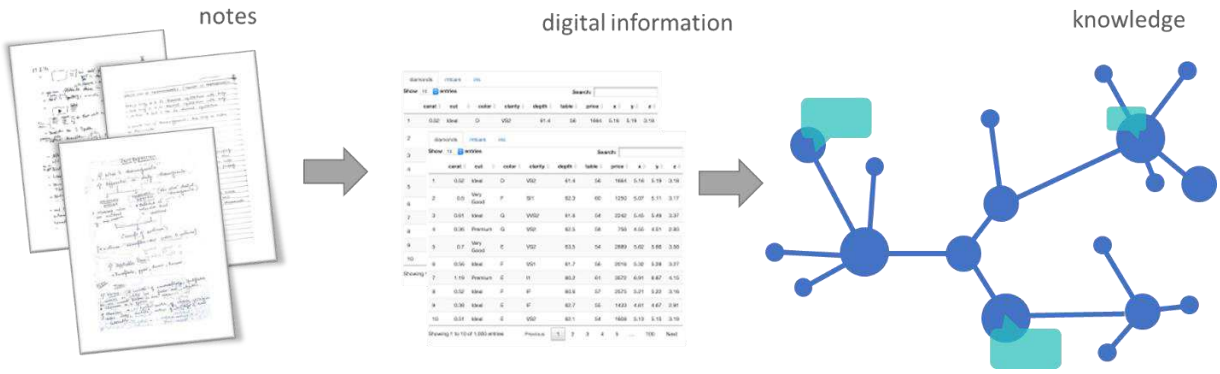

Figure 1. Becoming FAIR will drive science. Increasing the level of annotation at the source and tracking provenance using community standards will maximize data discoverability and reuse.

detect changes in thousands of different parameters responsible for complex plant behaviour. To interpret the resulting massive data sets, combination of various expertise from biology, chemistry, physics, mathematics and computer science.

Research data management services and infrastructures that facilitate the acquisition, processing, exchange and archival of research data sets 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.

Successful collaborative work and leveraging of data of different modalities – from many sources and experiments, and pre-processed or pre-analysed using a variety of algorithms – requires contextualization of the data according to the respective research objective. The FAIR Data and Linked Open Data principles provide crucial guidelines for any infrastructure receiving, processing and publishing research data [Figure 1]. While various consortia have made suggestions on best practices and processes towards fulfilling these principles, it is nevertheless always up to individual researchers' initiative to adhere to them. As a result, comprehensive information of the required quality for use by third parties is often only available in exceptional, rare cases.

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 and ontologies in plant science and beyond.

2. A robust, **federated research environment** for data computation and management covering the complete data lifecycle.

3. **Assistive mechanisms** ranging from data stewards to intelligent software services to build, link and maintain the complete research context during data acquisition, curation, analysis, and publication.

4. **Mechanisms for collaborative research** based on enrichment and automatized crosslinking of plant-research specific (meta)data to facilitate research context management.

5. A platform for data provenance and research sharing including a motivation and **credit system to foster the incentive** to democratize research data.

6. **Comprehensive training** to ensure data legacy through lectures, courses, workshops and summer schools and providing open training material.

7. A **central plant data HUB** for aggregating services and knowledge, generating a searchable compendium for research in plant biology.

DataPLANT provides an additional layer of services to provide facilities to complement existing generalist infrastructures and focuses on supporting and easing the processes of complete and meaningful research metadata context management which is often lacking or inadequate in fundamental plant sciences. In this manner, **we augment and complement existing services** in ways that go far beyond best practices currently used. DataPLANT ensures resulting well-annotated research data objects, ongoing qualification of data literacy for plant researchers, and an **integration of the plant research domain into the NFDI landscape**.

By the end of a five-year set up phase of DataPLANT we will have achieved:

- Knowledgeable researchers in the field, all students, PhDs, postdocs and PIs have a clear understanding of data management in the domain of plant research and are committed to produce perfectly reusable sets.

- There is a well-established first-point-of-contact in all relevant regards for researchers to learn about data management, relevant standards and research workflows in fundamental plant science. This hub is the established link between community members for further standard evolvements, it is the entry point to search for data in wide contexts. It is the link of plant research into the NFDI and the connector of the discipline specific data sets to the whole scientific community.

- There is a sustainable set of base level services available to the wider community to publish their data in a stable research data repository equipped with persistent identifiers.

- There is a viable long-term access service to past research contexts shared with other consortia in the NFDI and the wider scientific community.

- There is a search portal to make the provided research data findable according to the FAIR-principles.

2.2 Composition of the consortium and its embedding in the community of interest

DataPLANT provides user guidance and functionality to empower plant researchers in an open data world. Therefore, our two-fold strategy includes the development and implementation of an assisting data service infrastructure and the manifestation of the data generated among the experimental groups into the system in close collaboration. DataPLANT will lower the time and work spent on the user side to enrich data with metainformation. This added metainformation is necessary to render the data and to generate more value for the researchers and the community whilst decreasing efforts spend in data curation. The combination of federated intelligent software services and data management experts guaranty personalized assistive mechanisms tailored to the needs of the users. Close dialog with and early involvement of a domain-specific user community ensures relevance and usability of the system and safeguards the specificity and applicability of our requirements and data standards. Our domain specific user community is thematically coherent and brings together key actors from fundamental plant research including e.g. TRR 175-The Green Hub - Central Coordinator of Acclimation in Plants and the Cluster of Excellence on Plant Sciences (CEPLAS). The TRR 175 aims to discover how plants translate changes in light and temperature into cellular responses and identifies the molecular switches that are central to this. CEPLAS addresses the challenges for sustainable food production and ecosystem maintenance by fundamental research on complex plant traits of agronomic relevance that impact on yield and adaptation to limited resources. Additionally, our domain specific user community covers the fundamental plant research community being well distributed across Germany as well as in terms of data champions that generate the major amount of research data [Figure 2]. Methodologically, phenotyping platforms and high-throughput technologies such as mass spectrometry, next generation sequencing, spectroscopy, and imaging techniques, which are used in the user community of the consortium, cover the entire range of methods of modern plant research. However, in the field of plant research, manifold, diverse and time-consuming experiments need to be performed, accurately analysed and linked. Adequate metadata are required for the correct interpretation of the data in order to understand the mechanisms of life.

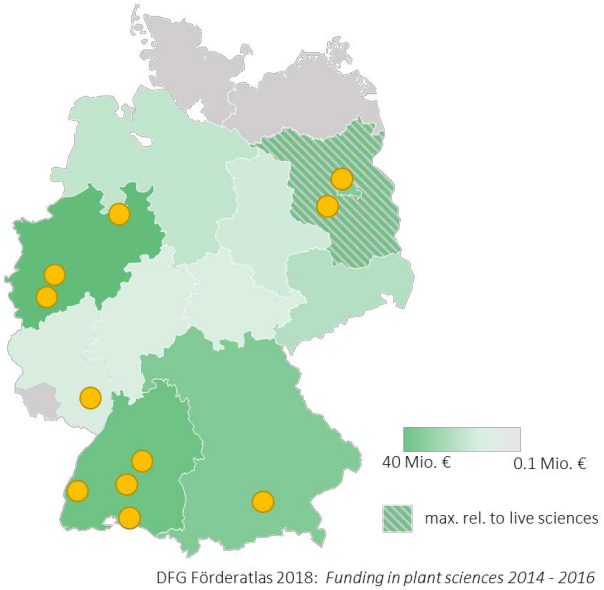

Figure 2 Distribution of the consortium across Germany. Federal sates are colored according to funding in plant sciences.

There is a large range of metadata, ontologies, data repositories and portals for software and computation, that in principle help analyse, interpret and share the research data. However, it became evident that researchers need practical support to cope with the fragmented and confusing landscape of resources to enable the democratization of research data. In addition to copyright and licensing concerns, finding appropriate repositories for deposition of data, was the need help to make data openly available according to the new report 'The State of Open Data 2018'[2]. Our consortium internal survey 'DataPLANT user survey 2019' aligns with the 'The State of Open Data 2018' report and emphasis the necessity for general support to enable FAIR data practice [Figure 3]. It started both from pre-existing research clusters in plant research at the applicant and partner institutions and from the efforts of the service providers to form a solid sustainable infrastructure to support relevant research groups at their home institutions. People involved in infrastructure projects like Galaxy[3], de.NBI or the Baden-Württemberg eScience initiative communicated with a broad range of potential partners, partially involving personal visits[4]. The use of existing project networks within the Galaxy, EOSC, ELIXIR, de.NBI cosmos and professional networks community forums were utilized over the course of the last year. These included the E-Science-Tage (March 2019 Heidelberg), SFB, Workshops, conferences, de.NBI Meetings (January 2019 Gatersleben, June 2019 Bremen), International Workshop on Science Gateways (June 2019 Ljubljana, Slovenia), CC Grid (May 2019 Larnaka, Cyprus), de.NBI Cloud User Meeting (September 2019 Heidelberg); Plant Acclimation Conference Irsee. Additionally, events hosted by the DFG or the Berlin meeting for cross-cutting topics (15.8.2019) were attended as well to foster a wider exchange between potential further stakeholders.

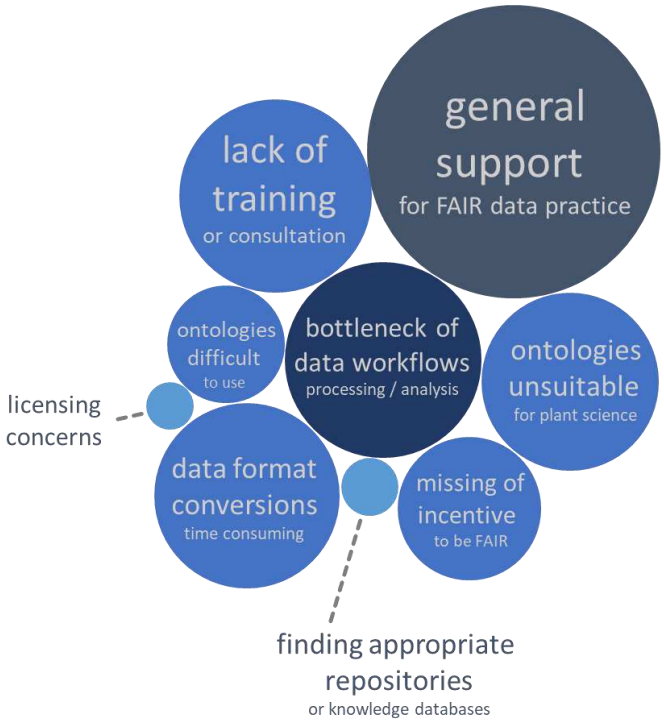

Figure 3 DataPLANT user survey 2019. The bubble size reflects the number of respondents that agree with the respective topic.

However, the DataPLANT survey allows us to prioritize and also evaluate our efforts in the future more formally. Most users have difficulties to use current ontologies for data annotation stating them to be impractical or not tailored to their respective research question. Most of the users

identified a bottleneck of computational workflows and bioinformatic support for analysing and processing of the data they are generating. By now many decentral resources got acquired and managed by individual groups, but they are local, unconnected to other groups or similar research fields. Much effort for keep-up and administration is duplicated and there is a lack of resources in long-term storage and access[5]. The situation of existing data sets regarding availability for reuse and verification is often problematic as more often data is just locally stored in the context of individual researchers[6]. Many researchers have the impression that they might fail the requirements of funders or publishers to properly provide access to data. Difficulties exist in sharing data due to patchy standardization. There is a widening contradiction between decade long (inter-)national functioning scientific cooperation and just beginning efforts required to run infrastructures to jointly work on data sets. Many research institutions are characterized by a waste of resources through lengthy and tedious processes to get a research project started: From acquiring the necessary resources, tedious workflow to find and prepare relevant data sets, to getting existing data interpreted and adapted to own workflows; this can be even worse for junior researchers regarding access to compute and storage resources. Up to now this is an effective barrier to the application of novel scientific workflows like machine learning and big data analysis. An additional very urgent concern seems to be the lack of training and/or consultation in general data literacy[7]. A more structurally related problem might be the missing incentive to be FAIR. Currently, there is simply not enough scholarly credit for well annotated research objects according to the FAIR principles compared to the value of classical journal publications. In addition, researchers typically consider data to be sensitive research outputs that can easily be misused or misinterpreted when taken out of context[8]. This raises the need for a research data management infrastructure to focus on assisting researchers to contextualize their research data according to the FAIR principles with a minimum of additional effort and skills chaperoning the full research cycle in modern plant biology.

It is necessary, that data management is driven by expert researchers in the field. However, the user community needs to be empowered to communicate their requirements by knowledge about the possibilities following the principle ‘application follows understanding’. It is essential for success to match and evaluate the theoretical requirements against real-world use cases. Therefore, an essential component of the DataPLANT strategy is the implementation of data stewards for bidirectional transparent communication. Data stewards are persons with high data literacy supporting data champions on site to custom fit RDM strategies and experimental work. In parallel, they are reinforced by our technical assistance services and infrastructure. Further, DataPLANT ensures transparent communication by instantiating an office as a single personal entry point that orchestrates contacts between providers and users. DataPLANT will also foster

various channels for direct communication within and across task areas like online team communication, social media and classical email service. Also, we aim to drive information exchange by training, organizing workshops and teaching in general. Regular surveys will help to further priorities the user requirements and they will be a necessary piece of our self-evaluation strategy. DataPLANT's organisational framework formed by a dedicated governance structure encourages and commits users to take an active role through general assembly, boards, and working groups. A more formal approach is having all participants committed themselves to direct communication by signing their particular DataPLANT letter of commitment to be part of this consortium.

Obviously, NFDI is a multilevel challenge and requires joining forces across disciplines. In order to cover the entire value creation chain, strong collaboration between plant scientists, data scientists, computer scientists, and organisational and infrastructure specialists are required. The DataPLANT consortium combines corresponding knowhow 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 utilizable for the fundamental plant research community. The expertise of the initially proposed consortia was on the one hand the specialization in plant scientist, data scientists, and computer scientists, and on the other hand knowledge and experience in computer scientists, and organisational and infrastructure. With the fusion to become DataPLANT, we will provide the glue between the disciplines and establish processes, communication and strategies to span and combine all necessary tasks required for research data management.

DataPLANT will provide a gateway to plant research data, ensuring open standards according to FAIR principles implementing a (meta)data standardization process based on international standards and rules that enables 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. Based on existing, open, and general metadata standards, we will be able to establish multiple example scenarios and templates covering most common workflow scenarios in the field. The considerable amount of research data generated by the consortium will allow us to train our data services to provide high quality templates and recommend adequate annotation information based on domain specific research-driven metadata for plant biology.

A broad range of experts complementing expertise and resources are participating in the DataPLANT consortium, that are required to establish a homogeneous interconnected infrastructure environment to enable modern plant research at the highest level, building on an

existing base infrastructure (BinAC, bwSFS, de.NBI cloud, bwCloud) and tailoring resources (ELIXIR, Galaxy) to the fundamental needs of plant researchers[4,9–11]. They bring in different strengths from both a scientific, organisational and provider perspective. The consortium will gain the expertise it needs to implement its work programme directly from the participants and the research and infrastructure network they are associated with. In special cases when certain expertise is not available e.g. legal questions (EU jurisdiction, country specifics), the consortium plans to hire experts in coordination with DFG and other NFDIs. Collaboration and harmonization with the other NFDIs will help DataPLANT to solve cross-cutting challenges.

DataPLANT goes beyond the simple user/consumers provider view and suggest a multi-layer model in which every layer is at the same time consumer and provider. These prosumers significantly profit from the surrounding and the exchange mediated by the NFDI. Several types of participants with distinct roles are present in the DataPLANT consortium and described in the following: (i) data champions; (ii) data and computer scientists, and (iii) organisational and infrastructure specialists.

(i) Data champions provide expert knowledge in plant research. They contribute significant expertise in next generation genome/transcriptome sequencing, high-throughput protein and metabolite analysis, mass spectrometry, advanced microscopic analyses, and general application of molecular biological tools.

Prof. Dr. Alisdair Fernie at Max-Planck-Institute of Molecular Plant Physiology, Central Metabolism, a data champion at the Max Planck Institute of Molecular Plant Physiology is focussing on metabolism and is able to analyse about 1000 metabolites (about half of them with known chemical structure). Here, the group uses this metabolomics platform to investigate natural variation of metabolism in e.g. domesticated and wild tomatoes, maize and beans. To understand this data the group is using genomic, transcriptomic and genetic data to identify underlying QTL. Furthermore, the group is investigating the metabolic response to abiotic stress where the group has unravelled important causal metabolites underlying UV response in *Arabidopsis*. The research methods in the group encompass metabolite profiling and targeted analyses using different technology platforms, as well as transcriptomics, genomics, computational approaches and GWAS[12].

The group of **Prof. Dr. Ute Krämer at Ruhr University Bochum, Molecular Genetics and Physiology of Plants**, a data champion at Ruhr University Bochum combines various genetic, genomic, population genomic and molecular physiology approaches in order to understand evolution, ecology and molecular mechanisms underlying evolutionary adaptations of plants to their local soil environment. Their present work focuses on the extremophile metal

hyperaccumulator species *Arabidopsis halleri* as a model organism. The group also studies how functional molecular networks operate in *A. thaliana* and its relatives and how their phenotypic outcome can be effectively modified. The research methods applied by the group include Genome sequencing, transcriptomics, population genomics, GWAS, quantitative genetics, molecular biology, molecular and classical physiology, ionomics, cell biology, biochemistry, statistical and field ecology.

The central research theme of **Prof. Dr. Dario Leister at Ludwig-Maximilians-University Munich, Department of Biology, Plant Sciences** is the molecular dissection of photosynthesis and of its interdependence of, and integration into, other cellular processes - within and outside chloroplasts. Photosynthesis-relevant cellular functions and their regulation within the organelle and in crosstalk to the nucleus are characterised by a combined approach, complementing genetic, biochemical, physiological and molecular-biological methodology with system biology approaches. The plastid-wide characterization of protein functions, in particular for photosynthesis, and of networks imposed on their regulation, will result into the redesign of the photosynthetic process by synthetic biology and experimental evolution. The applied research methods are Genetic (suppressor) screens in *A. thaliana* (including DNA-seq), quantitative biology (incl. Transcriptomics/RNA-Seq, proteomics and metabolomics (based on GC-MS and LC-MS), plant physiology and biochemistry, synthetic biology in *Synechocystis*.

The group of **Prof. Dr. Ralf Reski at University of Freiburg in Plant Biotechnology** has developed the moss *Physcomitrella patens* into a model organism for evolutionary developmental biology of early land plants (Funariaceae), systems biology and synthetic biology. Besides fundamental cell- and molecular biological research in the scope of evolutionary-developmental studies, the group's focus and specialization lies mainly on two research areas: (I) The continuous improvement of the moss genome and its structural and functional annotation, as well as high throughput analyses in comparative (phylo-)genomics and transcriptomics. (II) Analytical proteomics closely interlinked with the biotechnological utilization of *P. patens* for "molecular pharming", i.e. the production of therapeutic proteins in the moss bioreactor for an application in humans. *omics methods are commonly employed with high throughput analyses of NGS data from genomics, transcriptomics and proteomics. The DataPLANT NFDI would support the analysis and archiving of the *omics data and enable access to faster and more reproducible data analyses with standardized pipelines and workflows. Furthermore, an efficient access and visualization to genomic tracks of *Physcomitrella patens* and further moss genomes is highly desirable, e.g. via a well-defined web-service. This would increase the comparability of different natural ecotypes and mutant lines, but also the reusability of high throughput data regarding the model organism for the community.

The group of **Prof. Dr. Waltraud Schulze at University of Hohenheim, Plant Systems Biology**, is interested in regulatory processes at the plasma membrane in context with changing environmental conditions or nutrient deficiencies. Thereby, they focus on the regulation of nutrient transport by (receptor)kinases and undertake screens for ligands to yet uncharacterized receptors. The group studies dynamics of protein modifications (phosphorylation) and dynamic protein complexes and uses a combination of wet lab proteomic experiments with computational approaches to reconstruct signaling networks and predict their behavior.

The group of **Prof. Dr. Karl Schmid at the University of Hohenheim, Crop Biodiversity and Breeding Informatics investigates** the evolutionary history and environmental adaptation of crop plants. This is achieved by jointly modeling the phenotypic and genetic variation in combination with environmental data to identify environmental factors and genes relevant for crop adaptation. The main crops for investigation are major crops like maize and barley, and minor crops like amaranth and quinoa. The group uses DNA sequencing, in particular whole genome sequencing, field and laboratory phenotyping, and omics (e.g., transcriptomics) methods to characterize genetic and phenotypic variation. Subsequent analyses are based on population genetic and quantitative genetic methods like coalescent-based demographic modelling, selection tests and genome-wide association mapping (GWAS) to identify adaptive genes. Another line of research is the utilization of useful genetic variation in plant breeding by developing breeding methods that utilize genomic prediction and more recently machine learning approaches.

The research interest of **Jun. Prof. Dr. Sandra Schmöckel, at the University of Hohenheim, Physiology of Yield Stability** is to understand how some plants are able to grow in marginal environments and to find ways to make less tolerant plants grow better and maintain yield despite the presence of abiotic stresses. In the past she has been involved in a variety of topics, from characterization of transport proteins, mechanisms of signalling, genetics and genomics to field work, working with model organisms and crops. Her primary research methods are transcriptomics, genomics, metabolomics, physiology and phenotyping.

Prof. Dr. Andreas P. M. Weber at Heinrich Heine University Düsseldorf, Institute of Plant Biochemistry. Andreas Weber´s research program is centred on Molecular and Cellular Plant Physiology, in particular on cellular transport processes, plant genomics, and systems biology. He aims at understanding the molecular mechanisms underpinning C4 photosynthesis and its evolution, and employs systems approaches to understand the metabolism and ecophysiology of extremophilic algae. The Weber group has developed and applied tools for comparative (cross-species) transcriptomic approaches in a phylogenetic framework. Specifically, one of the first

studies reporting the application of next generation sequencing for transcriptome profiling (now known as mRNA-Seq) was conducted. In addition, the first quantitative comparison of related plant species at the transcriptomic level was performed, which led to the identification of a large number of candidate genes required for C4 photosynthesis. The Weber group is running the CEPLAS Plant Metabolism and Metabolomics Laboratory (various hyphenated mass spectrometry instruments). They generate substantial mass spectrometry data sets from plant, microbial, and animal cell metabolomics experiments and generate genomic data sets (Illumina, PacBio, Nanopore) from complex plant genomes (Brassicaceae, Asteracae, Portulacaceae, diverse algal genomes), genome sequencing projects in the context of C3/C4 and CAM photosynthesis. Plant biochemistry and physiology are getting applied including non-invasive phenotyping by reflectance spectrometry in combination with machine-learning approaches.

Dr. Sven Nahnsen is the director of the Quantitative Biology Center (QBiC) and research group leader in bioinformatics. His research focuses on FAIR data management and reproducible omics data processing. His research group initiated the internationally renowned nf-core project aiming scalable, automated and fully reproducible data analytics.

The **Quantitative Biology Center (QBiC) of the University of Tübingen** is the core facility for the central management and bioinformatics evaluation of large data sets in the life sciences. QBiC contributes already successfully developed data management components for these multidimensional data (genomics, transcriptomics, proteomics and metabolomics). These developments include intuitive user interfaces through which data can be annotated, shared and archived.

(ii) Data and computer scientists provide computational methods and expertise: They are experts in e.g. plant data standards, common workflows for omics data analysis, development and application of high-performance computational processing and analyses methods as well as data integration, visualization, analysis and interpretation as well as cross cutting omics data analysis.

Jun. Prof. Dr. Timo Mühlhaus at the Technical University of Kaiserslautern, Computational Systems Biology and his research group focuses on the application and development of computational methods to process and integrate quantitative biological data from modern high-throughput measurements in order to gain novel insights into biological responses to environment changes. The main challenge is the rigorous integration of different system level analyses and present knowledge into biological interpretable models. Therefore, we want to drive theory and technology forward with a combination of biological science, applied informatics and statistical approaches. Essential to this approach is the implementation of the methods developed and applied in our research in the form of (often application-specific) software packages used by

collaboration partners and the open source community. The development of a bioinformatic software library grants full control over the process from signal to information and final knowledge discovery. This sets the basis for customized solutions and methods to understand biological responses to environment changes on a systems level. Biological responses thereby represent a complex dynamic interplay between genes, proteins and metabolites. To understand these responses at the systems level, we need to study the structure and dynamics of cellular and organismal functions rather than the characteristics of isolated parts of a cell or an organism. Consequently, methods and models are required to capture this information accurately and efficiently.

Prof. Dr. Björn Usadel at the Forschungszentrum Jülich focuses on the analysis, visualization and interpretation of multi-omics data sets in plants with a focus on abiotic stress response and bioeconomical use. The approaches developed in his group range from data visualization, via outlier detection and statistical analysis to machine learning to predict traits or target genes in complex gene networks. Thus, his group is also annotating and curating large bodies of data one example being the MapMan ontology[13] tailored to visualization and statistical learning which is now available as a service to apply to all land plants. Due to his interests his group is spear-heading internationalization and standardization as well was open data efforts in plant omics and phenotyping data to make these data sets amenable to reusability and data mining procedures.

Prof. Dr. Rolf Backofen at Albert-Ludwigs University of Freiburg, Bioinformatics leads the chair for bioinformatics at the Technical Faculty. His research interests are the detection of RNA sequence/structure motifs, prediction and evaluation of alternative splice forms, investigation of RNA-protein and RNA-RNA and the description and detection of regulatory sequences. His group is one of the leading RNA bioinformatics groups with expertise in recognition, design and analysis of non-coding RNAs. He is leading the RNA Bioinformatic Center of the German Network for Bioinformatics Infrastructure and is ELIXIR Board Member.

Prof. Dr. Andrea Bräutigam at Bielefeld University in Computational Biology studies the evolution of complex plant traits using genomics, transcriptomics, proteomics, and metabolomics data. Comparative analyses of genomes and transcriptomes from algae over bryophytes and ferns to seed plants reveal the molecular underpinnings of observable traits and their evolution. For metabolic networks, stoichiometric and kinetic modelling are employed to mechanistically understand the role of particular genes. For regulatory networks, large scale transcriptome data arrays are analysed by correlation-based methods and with machine learning algorithms to identify the regulators of pathways relevant to yield, adaptation and acclimation to stress, and other production traits.

Prof. Dr. Stefan Deßloch at Technical University of Kaiserslautern, Chair for Heterogeneous Information Systems is a researcher in the field of database management and information systems. He will provide expertise in this domain. His main focus over the last years has been on data management in the cloud, data transformation languages and middleware, information integration and meta-data management, real-time data warehousing and analysis, as well as extensibility of database systems to provide search capabilities over structured and unstructured data. In the context of DataPLANT, Stefan is interested in addressing efficient and effective data management and query/retrieval/analysis support for experimental data and meta-data. The main focus would be on establishing architectures, data models, retrieval and analysis languages as well as efficient storage and access methods to support interactive, iterative and explorative query and analysis processes on large-scale experiment data and meta-data. Further fields of expertise are: Requirement analysis, architecture, language and system design, implementation and evaluation (including performance measurements).

Prof. Dr. Christoph Garth at Technical University of Kaiserslautern, Scientific Visualization, investigates methods for the processing, analysis, and visualization of very large datasets. His work draws heavily on formulating theoretically and mathematically motivated approaches, such as e.g. topological analysis, and translating them into applied techniques by developing corresponding efficient and scalable algorithms suitable for large-scale high-performance computing architectures. His recent work has emphasized complex data analysis and visualization workflows, where a variety of contributions were made towards composing complex workflows from simple building blocks, while still retaining efficiency and scalability through novel data management and parallelization strategies; in the context of DataPLANT, he will investigate these strategies for computational biology applications. Furthermore, a second focus of research has been on the visualization of data under uncertainties, such as e.g. ensembles, where novel techniques have been contributed to convey uncertainties in visual data depictions. Finally, an ongoing theme of research is the development of optimal analysis and visualization solutions for domain-specific problems across a wide range of domains, including among others computational biology, astrophysics, fluid mechanics, medical imaging. As a principal investigator in DFG IRTG 2057 “Physical Modeling for Virtual Manufacturing”, he considers these themes in the context of factory planning.

Dr. Björn Grüning at Albert-Ludwigs University of Freiburg, Bioinformatics, reproducible and accessible science, is leading the European Galaxy team, with over 8 years of experience working and developing with and for Galaxy. Björn is part of the German Network of Bioinformatic Infrastructure (de.NBI), the ELIXIR tools platform, the European Science Cloud (EOSC-life) and is responsible for the Freiburg part of the de.NBI cloud. Through this and the experience of

running the pan-European Galaxy server he has ample experience with managed systems and virtualized, cloud environments. As a core-member of the conda-forge, Bioconda[14] and BioContainers[15] community, Björn is steering the world-wide leading mechanisms for sustainable and reproducible software deployments. Moreover, his group is the driving force behind the Galaxy Training Network (http://training.galaxyproject.org) project to democratise and open training material, with a growing community that maintains more than 150 tutorials, ranging from Genome Annotation, Metabolomics, Imaging to Machine Learning. Next to the infrastructure work he is working in developing Omics pipeline with a strong focus on epigenetics. With deepTools[16] and HiCExplorer[17] he develops and maintains one of the most used software packages in this field. As technical coordinator of ELIXIR Germany and co-lead of the ELIXIR Galaxy community he will connect the DataPLANT consortium with de.NBI, ELIXIR and EOSC.

Prof. Dr. Dr. hc Edda Klipp at Humboldt- University of Berlin, Theoretical Biophysics, currently managing director of the Institute of Biology. The Klipp group carries out multi-disciplinary research projects to understand cellular organization, dynamics of cellular processes and stress response. Her group has long-standing experience in computational systems biology with focus on dynamic modelling of regulatory processes including signalling, cell cycle, metabolism, transcriptional regulation and growth control.**Prof. Dr. Eric Kemen at Tübingen University, Department of Microbial Interactions in Plant Ecosystems.** His research interest is to use high throughput sequencing methods (amplicon, metagenomics and whole genome sequencing) combined with modelling approaches to predict microbial communities that persist in nature and protect plants form pathogens by using field samples. The research focus of the group ranges from methods development for community network inference via genome sequencing and metagenomics to ecology and microbiology.

Prof. Dr. Heike Leitte at Technical University of Kaiserslautern, Visual Information Analytics, researches methods for the interactive visual exploration of scientific data. The work combines techniques from mathematical data analysis with visual interfaces to support the user in more efficient data analysis workflows. Applied mathematical analysis techniques include topological data analysis, information theory, classification and clustering. A major focus is the development of transparent analysis frameworks that communicate applied routines and potential errors in the data transformation process. This resulted into contributions towards the theoretic foundations of data visualization including work on visual saliency, error quantification, semantic analysis, and visualization quality analysis. Joint work was conducted with several application areas including 3D+T embryo and plant development from biology.

Prof. Dr. Klaus F. X. Mayer and Dr. Daniel Lang at Helmholtz Center Munich in Plant Genome and Systems Biology research evolution of plants and plant traits, through the comparative study of their genomes, gene families and gene regulatory mechanisms or networks. This entails comparing entire plant genomes and the encoded genes along the green tree of life to reconstruct ancient evolutionary events and traits, as well as comparing the genomes of many isolates or genotypes or populations within a single species or family to trace more recent changes and adaptations. The applied research areas are phylogenomics, systems biology, graph/network analysis, sequence analysis, ontology development and usage, text mining, data mining/machine learning, genome annotation/assembly using Bioinformatics, HPC/Grid applications.

(iii) Organisational and infrastructure specialists: They are experts in e.g. data standards, Open Access and/or Open Data provision, technical computing infrastructure organization, long-term data management and preservation, data retrieval as well as high performance and cloud computing. Due to their roles, they all have a long-standing track record in providing compute power and/or storage to user in a research data management context.

Dr. Dirk von Suchodoletz at Albert-Ludwigs University of Freiburg, Computer Center, is the head of the eScience Group since mid-2014 at the computer center of the University of Freiburg. He is an infrastructure specialist and his group is specialized in large scale research storage and compute systems as well as research data management. He is co-leading the Research Data Management group of the university and highly connected within the RDM working group of Baden Württemberg. His high-performance computing and cloud teams provide significant compute power in the de.NBI, bwCloud, bwForCluster NEMO and ATLAS HPC systems on well over 1200 servers. The teams implemented various innovative operation and deployment models easily integrating incoming requests of new research groups[18,19]. The systems are spanned by a high-speed network and supported by fast and bulk storage systems over 1 PByte. His specialized storage team organizes the setup and administration of a large-scale data management storage infrastructure federated with Tübingen to be used for higher level services data storage, versioning and repository services for research groups of different disciplines. Dirk's background is long-term preservation of digital objects and access to deprecated software and hardware stacks. He has extensive expertise in developing operation and business models for federated services[20]. He co-authored the governance structure within bwHPC and lead the development of the operating model of a cooperative, distributed-costs PC pool system with over 10 participating partners in Baden-Württemberg[21,22,22].

Dr. Jens Krüger at Eberhard-Karls University of Tübingen, Computer Center, leads the High Performance and Cloud Computing Group since 2017. He is responsible for various compute infrastructures at the University of Tübingen including the bwForCluster BinAC. The cluster serves among others the bioinformatics community in Baden Württemberg accompanied by the services of the state-wide bwHPC Competence Center for Bioinformatics lead by him. He is also operating the de.NBI Cloud Tübingen, a compute and storage environment for bioinformatics research as part of the German Network for Bioinformatics Infrastructure. Together with other de.NBI partners his group joined the European Open Science Cloud for Life Science (EOSC-life) earlier this year. His computer science related research focuses on sustainable science gateways and workflows. He has close ties into the corresponding community including the NSF-funded Science Gateway Institute. He is a member of the program committee and coorganizer of the International Workshop on Science Gateway. Together with expert from the Machine Learning community, he was the driving force behind the establishment of the ML Cloud Tübingen involving the Cluster of Excellence Machine Learning Tübingen, Tübingen AI and the Cyber Valley Initiative. The ML Cloud Tübingen is hosted and operated by the High Performance and Cloud Computing Group. His structural bioinformatics related research focuses on ion channels and their mechanisms of function. He is a member of the scientific advisory board of the Journal of Integrative Bioinformatics.

Olaf Brandt, Head of IT at Tübingen University Library is engaged in developing services for different library user communities on an organizational level. Technically the systems range from systems for specialists, bibliographic catalogs, metadata harvesting, specialized search engines, digitization-environments, publication systems, archiving-systems and integration with third party services. Olaf is involved in the development of Tübingen Campus Research Data Management services. He is long since engaged in digital preservation and preservation metadata. He is active in national (german digital preservation network nestor) and international developments. He is a former member of the PREMIS editorial Committee, a standardization effort, supported by the Library of Congress.

Dr. Marianne Dörr, University Librarian at Tübingen University is the director of the Tübingen University Library and the library system. As such she is engaged in digitisation, digital preservation, digital library services and services supporting science. On a national level she is a member of the Committee on Scientific Library Services and Information Systems (AWBI) of the German Research Foundation (DFG). She is a founding member of the Tübingen eScience-Center, which provides Research Data Management Services for the Campus and qualification and training measurements in data literacy and digital methods. Moreover, she is engaged in the development of the national Specialised Information Services Programme. Tübingen University

library provides for the discipline's theology, religious studies and criminology in Germany holistic services, e.g. special subject bibliographies via search engines, publication services on all levels, or support for research data management.

Petra Hätscher at University of Constance, is Director of the Communication, Information, Media Centre (KIM), which is the university's central service provider for IT and library services. The KIM is responsible for the coordination of Open Science policies of the University of Konstanz. Petra Hätscher is leading the project bw2FDM, a project on Research Data Management, funded by the Ministry of Science, Research and the Arts (MWK Baden-Württemberg) and a member of the bwHPC-S5 infrastructure support project. In addition, she was Program Director of several projects funded by the German Research Foundation in the area of Open Access and repositories: "Information platform open-access.net" (2006-2010), "Open access subject repositories" (2010-2012), "Move VRE" (2010-2012). She is a board member of the German Library Association. Furthermore, she was a member of the committee on "Scientific Library Services and Information Systems (US)" and chairwoman of the sub-committee "Electronic Publications" of the German Research Foundation (DFG).

Dr. Anja Oberländer is Head of Open Science at the Communication, Information, Media Centre at the University of Konstanz. In this role, she is responsible for all services, activities and projects regarding Open Access and Research Data at the University of Konstanz. Since 2007 she is coordinating open-access.net, the central German-speaking information platform on open access. Furthermore, Anja Oberländer is leading the program committee of the main German speaking open access conference “Open-Access-Tage”. She was project coordinator of the project "Open Access subject repositories" (2010-2012), funded by the German Research Foundation (DFG). She is also the project manager for OLH-DE, a project funded by the Federal Ministry of Education and Research, and responsible for the German National Open Access Desk in the European Commission’s OpenAIRE project.

Prof. Dr. Gerhard Schneider at Albert-Ludwigs University of Freiburg, Prorector, is head of the IT Center of the University of Freiburg. He served as CIO since 2009 and is currently Vice President for Digitalisation of the University. As VP he not only tries to introduce new IT Support concepts in order to make professional IT Support affordable to researchers, but also pushes the ideas of research data management across all disciplines. For many years, he also served on various committees of the DFG and DFN. He is one of the authors of the State HPC concepts. Already in 2003, he founded the New Media Center, a Cooperation with the university Library, to focus on service issues and to avoid doubling of structures. As a result, E-Learning is a strong

asset of the university. He is currently head of ALWR, the IT commission of the State University Rector‘s Assembly.

Dr. Inga Scheler at Technical University of Kaiserslautern, Computer Center (RHRK), is the Vice Director of the Computing Center (RHRK) at University of Kaiserslautern She holds a PhD in Computer Science from University of Kaiserslautern (2008) and has about 20 years experience in the research fields information visualization and data analysis as well as basic IT-infrastructure.

Prof. Dr. Thomas Walter researches Information Services at University of Tübingen. He is head of the IT centre of University of Tübingen and Chief Information Officer (CIO). He established High Performance Computing there and is responsible for research data management. Together with Dr. Marianne Dörr he is a founding member of the eScience Center. He’s author of bwDATA 2013-2014 and 2015-2019, the concepts of Baden-Württenberg’s Universities to handle scientific data, and of *Umsetzungskonzept der Universitäten des Landes Baden-Württemberg für das High Performance Computing (HPC), Data Intensive Computing (DIC) und Large Scale Scientific Data Management (LS2DM).* 2012 Thomas established Bachelor of Science Medical Informatics studies at University of Tübingen, followed 4 years later by Master of Science in Medical Informatics. Since 2014 Thomas is chairman of Hochschul Informations System (HIS) at Hannover.

The **IT center of the University of Tübingen (ZDV)** coordinates all IT-related activities at the university. Large-scale compute and storage resources, including the bwHPC Cluster BinAC, the de.NBI Cloud Tübingen and the ML Cloud Tübingen are operated. Further, it has a long-standing expertise in hosting data and providing storage. The ZDV is involved in multiple research data management activities, either as participant, through infrastructure hosting or as resource provider. These activities comprise among others the eScience Center Tübingen, the QBIC, the CiTAR project or the Campos project. The **computer center of the University of Kaiserslautern (RHRK)** operates in a federated setup in the state of Rhineland-Palatinate. RHRK excels at providing customer-oriented, tailored data management solutions and services. For example, all data processing for the TR-SFB 175 "The Green Hub" among TU KL, HU Berlin, and LMU Munich is centralized at RHRK through a tailored data and compute infrastructure. Here, RHRK provides both hardware and software services. RHRK brings substantial proficiency in operating in federated environments that provide transparent access to de-localized storage and compute services. Beyond providing the expertise in setting up such environments, RHRK will also contribute local compute and storage capacities into the DataPLANT environment, e.g. the HPC cluster Elwetritsch and a Microsoft Cloud environment that is currently being prepared for roll-out. In operating services on behalf of academic customers, RHRK has gathered significant

experience in development and long-term operation of customer-friendly usage models. Developing such models is a central, major challenge facing all NFDI consortia, and RHRK would be happy to contribute in this area.

The **computer center of the University of Freiburg - complemented by the professorship in Communication Systems of the Technical Faculty** - has extensive expertise in long-term storage and access to digital objects and research contexts. It was involved in several large scale European and national and state-wide research projects and supports a world-wide unique access services to past computer environments (Emulation-as-a-Service). Through the long-standing participation in large scale federated research infrastructure projects of universities in the state of Baden-Württemberg it gained experience in creation, operation and governance of cooperative infrastructures. The computer centre hosts all crucial university infrastructures in georedundant server room locations and maintains redundant high-speed uplinks to the state-wide research data network. It operates the major IT systems for both the university administration and the faculties as well as various large-scale scientific compute and storage systems[23]. It is the leading entity in the university to develop and implement the research data management strategy for the whole organisation.

The KIM at University of Constance in behalf of the university is one of the pioneers and unremitting proponents of Open Science in Germany. The Communication, Information, Media Centre (KIM) is the university's central service provider for IT and library services. In 2012, the university published their Open Access Policy, declaring Open Access to scientific publications to be the guiding principle for their scientific publication strategy. In 2018, the university's Senate passed the Research Data Management Policy, thus confirming the university's commitment to the "responsible handling of research data to be a foundation for transparent and efficient research". KIM is part of the research data management-project "bw2FDM", a joint venture of universities of the federal state of Baden-Wuerttemberg to survey how scientific communities deal with research data and to assess their need for research data management support, infrastructure and services. Furthermore, the project operates and develops the website "forschungsdaten.info", which is an information service for researchers providing information material on the topic of research data management (RDM) in German. Besides KIM is a member of the Science Data Centre BioDATEN and the project Movebank 2.0. KIM supports both projects in infrastructural fields. The university is leading in the field of Open Access in Germany since it has got the highest Open Access rate of all universities in Germany.

The **Galaxy team, partial at Albert-Ludwigs University of Freiburg, Bioinformatics,** is maintaining the European Galaxy server (https://usegalaxy.eu server), with an emphasis on

reproducibility and accessibility to tools, workflows and data in the Life sciences. The aim of the Galaxy project and the Galaxy team in Freiburg is to enable scientist to perform bioinformatics and biostatistical analyses in a reproducible and transparent way. Galaxy provides the entrance portal for the scientist and makes it possible to reproduce this workflow in containers on computing infrastructures (HPC or Cloud). For this, the European Galaxy Server will be offered. The Freiburg Galaxy Team (10 employees) contributes many years of experience in the field of surveys and analysis of data from genomics, transcriptomics, epigenetics, proteomics, metabolomics and metagenomics. With over 2.000 publicly accessible tools on the European Galaxy server it serves almost all communities of the life sciences. The Galaxy Framework stores all provenances that belong to the reproduction of an analysis. In the end, the framework itself as the execution layer of the Tool Container must also be preserved. The collaboration of the data champions with the workflow developers and long-term access and reproducibility experts will provide the necessary input to ensure this. The Freiburg Galaxy team is part of GOBLET, the ELIXIR Training platform, the Galaxy Training Network (GTN)[24] and has certified Carpentries training instructors. The team has provided dozens (https://usegalaxy.eu/freiburg/events) of Galaxy training courses world-wide for developers, admins and scientists. Capacity building is done by Train-the-Trainer workshops and various Mentoring programs, e.g. together with Mozilla. DataPLANT and the Galaxy Training Network (GTN) facilitate the transfer of knowledge and support of the relevant tools and workflows of the community in workshops and supplemented with of e-learning material and online tutorials. Galaxy provides a wide range of sample trainings and demonstration material to be used in such qualifications.

2.3 The consortium within the NFDI

State of the art life sciences as in the field of plant research are characterized by a couple of phenomena. The digitalization of scientific workflows significantly changed the methods of progress and knowledge gathering as well as the workplace of each researcher. Digital means of communication allow for a much faster exchange and collaboration between research groups and greatly increase the speed of scholarly communication. But just like in the analog world of the past, the main focus of researchers is still on traditional output in the form of papers and articles. Related data and research contexts too often are neglected, which results in a challenge to reproduce experimental findings. This is due to short term orientation in research as only novel insights will be published and honoured by the scientific community. The proof of older results is rarely acknowledged and thus scarcely done. The NFDI and the DataPLANT consortium acknowledge the actual state and strife for a paradigm shift in digital workflows, data management and publication of results. In the environment of modern science which is both competitive and cooperative, solely self-regulating these challenges fails and the necessary infrastructures for

sustainable long-term access and publication of data are missing or scattered. The competition and organisational challenge hinders the plant researchers from working on a common infrastructure for data management. DataPLANT within the NFDI will play multiple roles: It will primarily provide a designated scientific community with its domain specific requirements and expectations. It will cooperate in the general NFDI by contributing to cross-cutting topics (e.g. participating in the Berlin Declaration), providing and consuming services open to a wider community. DataPLANT will promote the cultural change towards a widened concept of crediting research through well annotated data and workflow publication, a new cooperation model based on infrastructure and standardization. It will recalibrate the balance between competition and cooperation. DataPLANT aims at maintaining the competition for advance in the field of plant research but aims at the paradigm change to do so by striving to publish well-annotated data in a reusable way in a wider NFDI context beyond our own focus group. This would allow the NFDI to establish a self-sustaining cycle of data centric insight: Researchers get rewarded for well-annotated research objects and gain credits for research, motivating them to increase volume and quality.

DataPLANT works towards a collaborative governance and common framework of services. The governance concept anticipates a transparent, user-centric implementation of organisational structures. DataPLANT brings in the infrastructure provider's long-term experience in federated infrastructures, and expertise in setting up cooperation 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 as well as infrastructure and service providers. DataPLANT data stewards as experts operating between the researchers and the infrastructure/services play a crucial role in reaching out to the community and allow for an aggregation of concepts, data and expertise. Both approaches could serve as an integral building block for other NFDIs. DataPLANT is closely collaborating at the national and international level within the plant/bioinformatics community, spanning a collaborative meta NFDI network. DataPLANT will use its roots within the participating universities to train and raise awareness about modern data management in general and specific concepts of DataPLANT. Through the extension of curricula regarding data management and data analytics DataPLANT will extend the recruitment pool of qualified personnel. Further, the international network of scientists can help to fill staffing gaps.

Cross-cutting topics. All consortia present at the Berlin meeting mid-August agreed that strong cooperation and communication among NFDI consortia is essential for building a meaningful research data infrastructure for Germany in an international context. The cross-cutting topics shall be addressed by several consortia in inter-consortium working groups. Being a

member/contributor to the “Berlin Declaration of handling cross-cutting topics”[25], DataPLANT will get involved and contribute to the agreed-upon topics. DataPLANT will work together with other consortia on the common vision of the NFDI including long-term foresight and common strategic planning. It coordinates policy advice, consultation and outreach with the other NFDI consortia. The same applies for human resource management, recruitment and development. It will bring in its activities regarding cultural change on reputation, publication, funding policies and novel credit for research systems. In particular, DataPLANT helps to increase the international visibility of the NFDI through its active international networking. A core objective of DataPLANT mirrored in the first task area is the standardization of metadata and harmonisation of services. Ontologies are the core for data description and understanding by the designated community and informed third party and a basis for metadata definition and subsequently the findability of data sets. Typically, harmonisation is required and the procedures in each community need to be moderated. This also involves terminologies, terminology management and services. Other consortia like NFDI4MSE might offer suitable processes.

Collaborative governance and general NFDI framework. A national infrastructure providing services to the whole scientific community requires appropriate governance and structures to balance the needs and expectations of all involved parties. Clear governance structures are a key prerequisite to ensure sustainable operations of a distributed infrastructure like the NFDI. Therefore, a major challenge will be the identification of an appropriate legal entity which serves the interests of the consortia and service-providing host institutions. DataPLANT will contribute state-wide service federation and cooperation from its members’ long experience[20,26,27]. Discussions are needed about sustainability, operating, cost-covering and legal models for the coordination bodies/offices of the NFDI and consortia. For the NFDI as a whole, decision-making powers and structures must be coordinated, agreed upon and put into an appropriate organisational structure. The work of the consortia in relation to cross-cutting topics and the NFDI network should be coordinated and planned. It would make sense to define a comprehensive plan of activities with milestones for the NFDI forum as well. Quality assurance measures of the implemented structures should be agreed upon and implemented. Financial resource flows and the distribution of funds are to be organised. A challenge for the NFDI as a whole is that very different disciplines, each with their specific culture, come together and thus the network is more like a 'team of teams' than a hierarchically controllable entity. In addition, there is a need to lay down rules for cooperation without, however, falling into over-regulation. The NFDI is to be understood as an evolving process, which is why not all aspects can be controlled down to the last detail by rules. Nevertheless, in order to enforce the jointly agreed rules, possible sanctions should be defined.

A forum of the consortia should be set up to steer the development of the NFDI, clarify and organise the processing of cross-cutting topics and cross-disciplinary coordination and, where appropriate, standardisation. In addition to organisational issues, this forum should also address common issues of financing, personnel management and development, and overarching continuing qualification. One measure is to cover training and qualification activities on aspects of management of the consortia and their networks. A further measure is the creation of platforms in which a guided and structured exchange on challenges and problems of research management, but also on procedures and 'good practice' is initiated and maintained. A supportive measure of a completely different nature is to provide the (co-)speakers of the partial NFDIs with tools that help them to further develop the individual consortia and draw on qualified external support for certain tasks if necessary.

Community (User) involvement. User involvement and motivation is essential for a successful NFDI. Novel approaches to motivate users for appropriate data management need to be discussed and evaluated across all consortia. DataPLANT implements a user driven adaptive development of the consortium and will deploy the same spirit in developing cross-domain use cases. It participates in the coordination of teaching and training as agreed upon with other consortia such as NFDI4MSE, undergraduate and graduate education (curricula) and professional development. We implement the necessary organisational structures e.g. through working groups to join the dynamic development of NFDI and its (meta)data standards.

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.

FAIR data compliance -a core component in DataPLANT- surely is a key objective for all NFDI consortia and will ultimately lead to interoperability across domains. Automatization of FAIR compliance that minimizes the user effort might be extendable to other NFDI consortia. We are eager to discuss and contribute towards a NFDI-spanning solution. Developing common views about the quality assurance of NFDI repositories and services (and the data provided by them) is

a further important cross-cutting topic, which will also be of relevance for the NFDI governance, the necessary cultural change, and user engagement.

Provenance. For all scientific communities, the origin of data and its modification over different workflows are factors 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 (filter) processes. Many findings will be directly transferable to other consortia as well. This makes it possible to transparently track any possible change of a data set, 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.

Legal and ethical aspects: Legal aspects such as licensing of data and software, intellectual property rights, data protection and privacy are of utmost importance for communities dealing with sensitive data, but they cannot be ignored by other communities either. The same holds for ethical aspects. All these topics would greatly benefit from a consolidated approach of the data users and data providers within NFDI. Likewise, as with many other cross-cutting NFDI topics, most legal and ethical aspects need also to be considered in the international context: Research data often stems from international collaborations or is shared with international colleagues. A specific concern with relevance for the development of a sustained NFDI governance is to clarify parameters for the commercial use of data and the potential commercialisation of data. DataPLANT identified a general need by the plant community in legal support and thus plans a person to fill that gap. This role will both coordinate with the other consortia on common problems and contribute to the NFDI as a whole.

Sensitive and especially person-related data to be used for research adds particular challenges, both on the ethical and the legal side. The special protection this type of data requires to maintain a person's privacy and to fulfil the strict legal requirements imply a larger effort on the technical side in order to ensure proper data protection, but also requires ethical considerations as well as a legal framework providing transparency and legal certainty on the use of the data for both data providers and data users. We do not expect GDPR-related issues with plant data in our community. Nevertheless, ownership and responsibility play a role. DataPLANT might need to deal with dual-use challenges and the adherence to the Nagoya declaration. In rare cases, special

data protection measures might be required. We will learn from the respective measure from the medical field.

Technical infrastructure and concepts: Research data management inevitably has implications on a rich set of topics concerning technological implementations and (computer science) concepts. To allow for cross-disciplinary access and reuse of research data within the NFDI, some level of standardisation and harmonisation is required for several metadata and data properties. Several consortia share the view of a Research Data Commons (RDC) as an overarching virtual expandable infrastructure to leverage user involvement and collaborative data-driven research. This includes for example joint cloud services, access to computing power and collaborative workspaces, and a common authentication and authorisation infrastructure (AAI)[28]. The Research Data Council calls for a common strategy for interacting with the existing large-scale compute and data infrastructures in Germany and the need for harmonisation among these centres. The DataPLANT consortium brings in significant contributions regarding technical infrastructure and operation concepts and expertise. The infrastructure is partially shared with other NFDI consortia run in a federated operational model. The scalability is an inherent feature of the DataPLANT concept expressed through a federated infrastructure. The handling and orchestration of scientific workflows is done through Galaxy which offers interfaces between data providers and users across research domains. We will build upon existing identity federations like ELIXIR AAI and will contribute to a NFDI-wide standard in this regard.

Preservation of the research context. In pretty much every field, preserving just the data objects risks losing access to the research context, and thus, eventually the ability for data interpretation and data reuse. Hence, data, data-processing software and sometimes even base-level technology stacks need to be considered in a joint context. General considerations for long-term (ten years and more) reuse, validation and reproduction of research outputs is still in its infancy. The DataPLANT 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 with national and international consortia, as well as guidelines and infrastructure to manage and preserve software dependencies which should be made available to all NFDI initiatives. Still, 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 ensuring meaningful access to data and reuse requires constant attention and development. In order to ensure FAIR data principles, especially 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.

An adjacent field is quality management and assurance including the certification of services. Criteria are to be developed and agreed upon for data, software and services. DataPLANT will evaluate and evolve such criteria in a work package and would be willing to contribute to formal certification processes for NFDI service offerings.

Specific coordination and interaction with other consortia. DataPLANT combines corresponding expertise of the initially proposed BioDATEN4NFDI and the DaPLUS consortia. In general, DataPLANT will provide a gateway to plant research data and metadata, ensuring open standards according to FAIR principles. 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 which is also safeguarded by one of DataPLANT's co-speakers being responsible for standards in NFDI4Agri. 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 NFDI4Neuro to improve generic data workflow management. On the infrastructure level cooperation with other service provides like the RHRK and other research institutions computer centres 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[7]. 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 and FAIRmat, the exchange of basic RDM and molecule-specific training materials between the initiatives is planned. DataPLANT will learn from consortia like NFDI4Chem on the

concept of Electronic Lab Notebooks and the embedding of them into digital workflows. It wishes to create a common base standard with other consortia to be useable for the plant research community.

NFDI4MSE and DataPLANT share institutional and personal ties and the same spirit, trying to support modern research workflows and to foster a cultural change in their research domains. NFDI4MSE acknowledges the advanced starting point that DataPLANT can found its efforts upon, already having a steadily growing platform at its disposal through the Galaxy Project. NFDI4MSE looks forward to better understanding their technical infrastructure and transfer lessons learned into the shape of the MSE data space. Here, especially Galaxy's modular toolbox for data processing and computing, allowing for a flexible integration of newly added tools promises an interesting starting point for the development of comprehensive interfaces, while at the same time ensuring the necessary adaptability. At the same time, NFDI4MSE has advanced concepts at its disposal for a comprehensive approach to infrastructure development, teaching and outreach, some of which DataPLANT could profit from. This includes, e.g., the thematic fields of persistent identifiers[29], decentralised raw data access points, authentication infrastructures, but also the more organisationally relevant approaches to education or business and incentive models. NFDI4MSE and DataPLANT will discuss and advance these topics on the general NFDI platform. Obviously, any such harmonization effort in these crosscutting fields lays groundwork for the future integration of different NFDIs within the shared NFDI association. The vision is, as integration is inevitable, efforts should start to integrate right from the beginning.

Galaxy has recently gained support to analyse molecular reactions and interactions[30] with the aim to study biomolecules. However, the underlying tools like GROMACS[31] support much more use-cases, e.g. material science. Therefore, we would like to work together with FAIRmat to connect our workflows to the NOMAD repository. A "Galaxy data-source" is a convenient integration that redirects a user from Galaxy to the NOMAD user-interface, lets the user filter and select datasets and streams these data back to Galaxy without the need to store data on the user’s computer. Vice versa, Galaxy could also upload results to NOMAD and treat it as persistent data storage after processing complex workflows on some initial dataset.

Beside its activities in DataPLANT, the computer center of the University Tübingen will be engaged in GHGA and NFDI4Earth. The computer center in Freiburg was asked to support NFDI-Neuro and the particle physics consortium PAHN-PaN with infrastructural components e.g. using the HPC cluster NEMO[23]. The Forschungszentrum Jülich IBG-4 participates in NFDI4Agri.

DataPLANT recognizes the NFDI as an ultimate chance to coordinate cooperation in data management and common services for long-term access to research contexts. The NFDI can

help to overcome fragmentation of efforts and foster the necessary cultural change. It should put itself into the research lifecycle and should require endorsement by funding agencies and publishers. Especially the latter should accept Open Data and link to publicly available and sustainable data repositories. DataPLANT expects a clear commitment on that change by offering all stakeholders clear incentives for participating in high quality research data management as envisioned by our consortium. Besides institutional funding, viable and long-term funding streams are required in order to offer a permanent perspective for data stewards and to allow a regular update and adaptation of the necessary infrastructure. The refinancing of compute and storage systems needs to be coordinated on a wider level. Special services such as for maintaining reproducible execution environments, which might be necessary for long-term access to older data sets, should have a clear commitment and be coordinated on a national level.

2.4 International networking

The consortium comprises major players in fundamental plant research and it is well integrated into the Europe and international research and IT landscapes. On the IT and infrastructure side, DataPLANT is embedded deeply into the European landscape, on the one hand side the consortium comprises several members of the German ELIXIR node (i.e. the pan-European infrastructure for biological information) in the areas of cloud computing, tools/ workflows and plant bioinformatics and data analysis. Members of DataPLANT are technical coordinators in ELIXIR, co-leads of the ELIXIR Galaxy community and part of the ELIXIR Tools platform as well as the Plant community. Freiburg is also part of the European Open Science Cloud (EOSC) and Jülich is a member of the EOSC-LIFE consortium. EOSC is a European Commission project to provide a public data repository which conforms to open science values and aligns well with the aims of DataPLANT. Consequently, the European Galaxy server is part of the EOSC marketplace. Beyond ELIXIR, there are also tight links to the European Bioinformatics Institute (EBI) e.g. in the area of Plant genomes (ENSEMBL Plants).

On the data analysis and storage side, the partners contribute as members to e.g. EOSC-LIFE providing data integration across life science infrastructures and are involved in the European Plant Phenotyping Network (EPPN), the phenotyping activities of the ESFRI listed project EMPHASIS, the COST action (CA16212) on “Impact of Nuclear Domains On Gene Expression and Plant Traits” and International Plant Phenotyping Networks (IPPN) where DataPLANT members contribute to IT and storage and in the Research Data Alliance (RDA). Furthermore, DataPLANT is involved in the effort to unlock diversity (DivSeek) and the International Workshops on Sciences Gateways (IWSG) and is collaborating with Cyverse (formerly iPLANT), which provides data and services chiefly for US researchers, and Cyverse UK.

Building on these interactions and the link to the plant phenotyping community, DataPLANT also participates to the “Minimal Information about a Plant Phenotyping Experiment” (MIAPPE)[32] consortium currently as a steering community member and pushes the standardization efforts of the plant community. On plant specific infrastructures and international consortia, members of DataPLANT collaborate with a plethora of international database and plant infrastructures such as Araport, the Canadian BAR, the Singaporean CoNekT/Planet resource, Belgium PLAZA, the Australian SUBA and multiple plant genome and transcriptome sequencing and analysis consortia (e.g. IGWSC, 10+ wheat, Chara, Cuscuta, Physcomitrella, potato pangenome etc.). Sustainability of projects can only be reached if an international community can be built around it, so that local changes in research directives and/or temporary lack of funding can be compensated. DataPLANT has an impressive track-record in building sustainable communities and growing them over the years to multiple thousands of contributors world-wide. As part of the Galaxy steering committee, the conda-forge and Bioconda[14] core team and as co-founder of BioContainers[33], DataPLANT has members that are driving the international scientific research infrastructure since over a decade. It is estimated that there are over 200.000 Galaxy users world-wide. There are more than 7000 Bioconda packages and Containers available that have been downloaded more than 15 Million times. Furthermore, many of the participants are journal editors and partners advising on metadata and omics standards which helps in setting up standards and procedures through journal recommendation for authors. The RDM research team is internationally engaged in setting up U.S. national software preservation infrastructure through the Software Preservation Network, the Emulation as a Service Infrastructure (EaaSI) initiative and through EaaSI with the PresQT project. On the teaching side, the University of Freiburg - under the umbrella of Eucor – The European Campus - is engaged in trinational cooperation with the top universities on the Upper Rhine in Basel, Strasbourg, and Karlsruhe as well as with the university in Mulhouse to strengthen cross-border activities in research, teaching, and transfer. One of the showcases is a joint biotech master programme, where the group of Ralf Reski is co-teaching the Plant Biotechnology module both as a lecture and a practical hands on course.

2.5 Organisational structure and viability

The DataPLANT organisational structure and governance is set up to foster efficient communication and deal with at least three domains:

- Internal project governance and financial operations (grant money, additional resources, etc.), and assignment of data stewards to individual researchers and groups. The internal governance defines the rights and obligations between the NFDI participants and the steering body presented by the DataPLANT boards.

- Interaction between providers and users; support an active role of the scientific community, coordinate change process aimed at technical, organisational or structural enhancements.

- Inter-NFDI coordination to advance the cross-cutting topics, foster cooperation and evolvement of the organisational structures.

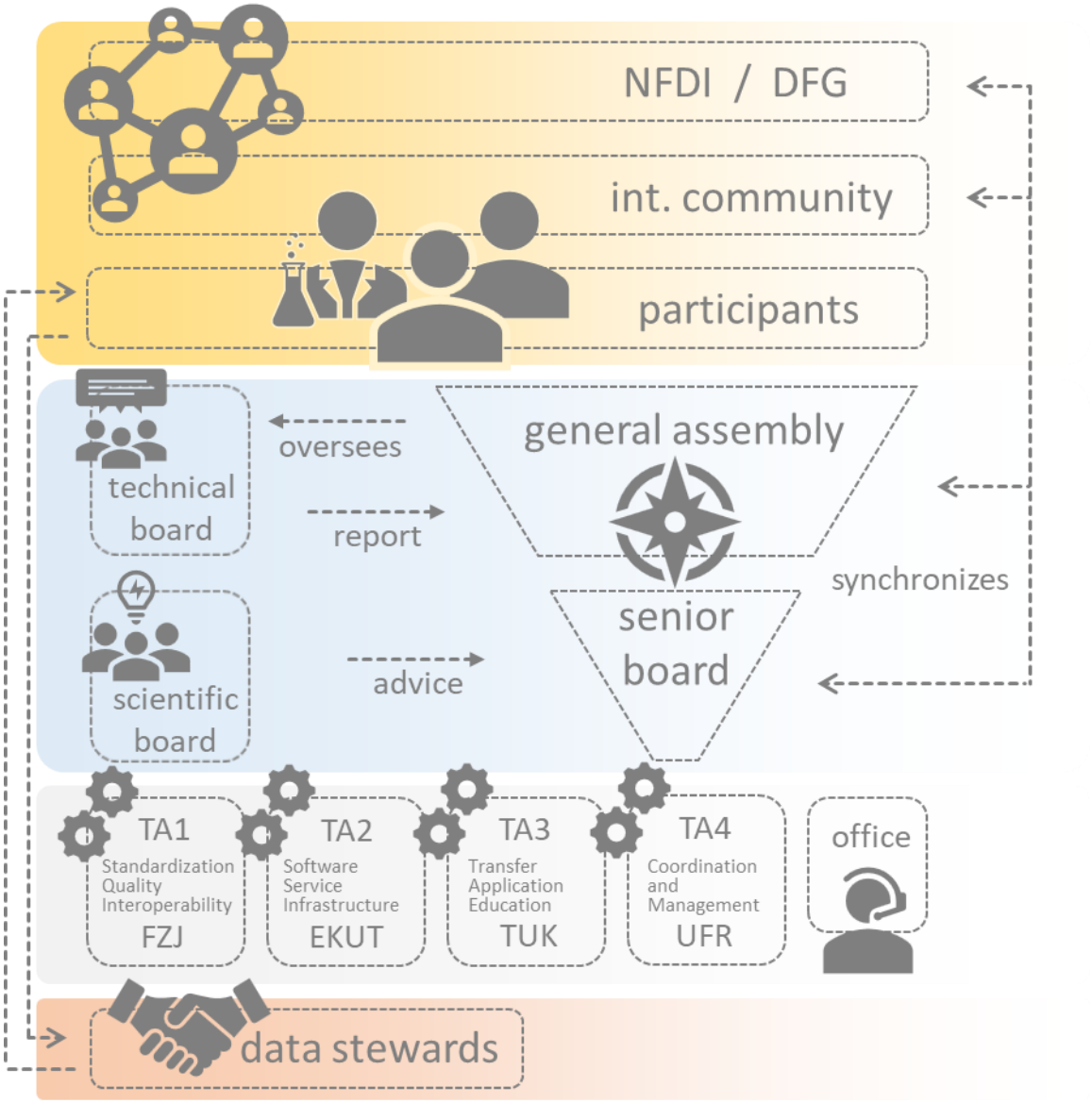

Figure 4: Overview of the organisational structure of DataPLANT

Organisational structure. In DataPLANT, three groups of stakeholders are present: The DataPLANT scientific community, the service providers, and the system and services developers [*Figure 4*]. The users indicate what kind of base-level services they require to conduct their research, fulfill the requirements regarding scientific code of conduct, and to run their own (high-level) services. The interests and objectives of the three groups and stakeholders have to be

taken into consideration and balanced against each other. To this end, advisory committees were created in the form of the scientific and technical boards. The technical board consists of technical experts and moderates the various infrastructure requirements and data service offerings and decides on resource distribution/allocation, future development of new offerings or the deprovisioning of deprecated services. The scientific board consists mostly of plant scientists and computer scientists developing services and advises on technical needs and develops foresight processes. The boards take the input from both the general assembly and the senior management board and outside requests (e.g. from other consortia or the general NFDI) brought in through the DataPLANT office. These bodies will take care of the strategy and standards development and suggest consortium members as experts for the relevant working groups. The data champions play a special role as they tightly interact with the task area managers to shape and adapt the development agenda. The data champions are "super users" which are presented both through the general assembly and directly through the work packages in the task areas. They will participate as well in a special governance body - a working group to evaluate data sets, decide about obsolescence, define decision and quality assurance metrics and identify strategy gaps. Primarily the speakers and the senior management board will facilitate the inter NFDI exchange and coordination with the NFDI board. They collect requests and input to the overall NFDI development in the general assembly meetings and scientific board sessions and bring this to the general level. They coordinate the input from DataPLANT to the cross-cutting topics. The internal governance defines the rights and obligations between the stakeholders.

Moderating decisions. Both technical and scientific boards moderate requests, which were not decided within the respective body to the high-level decision-making committee, *the senior management board*. The *senior management board* consists of all *co-speakers* of DataPLANT and reports to the general assembly. Whilst the senior management board strives to make unanimous decisions, it will employ a simple majority vote principle where in case of a tie vote the coordinator decides. The *office* handles the everyday business on behalf of the management board and the disbursement of funds of the sub NFDI. It is the first point of contact for the general NFDI processes, the facilitation of the inter NFDI exchange and the DataPLANT community (both participants and new users). It acts on behalf of the management board in between meetings and buffers incoming requests. The office takes care of the financial affairs of DataPLANT and reports annually to the general assembly and back to the grant provider. Furthermore, it handles inter-NFDI affairs and supports the coordination of cross cutting topics. The development of the governance structures was initiated during the consortium setup process and with the involvement of all relevant parties. The NFDI-wide governance structures encompass all affected types of resources, taking strategic decisions. Additionally, dedicated and modular governance

structures were established which directly address the needs of individual researchers. The DataPLANT governance and coordination is designed to be open for adaptation; the NFDI governance -which is not fixed yet- was identified as a cross-cutting topic relevant for other scientific communities as well. The suggested structures will be put into a regular reviewing scheme by the general assembly and the senior management board. Both inputs from the DataPLANT community as well as from the NFDI in general will be considered. Data stewards are the facilitators between users, providers and developers and are represented by a further internal steering body, overseen by the senior management board.

Rules and commitment. DataPLANT proposes rules for cooperation, distinguishing three levels: A) The level of results - here it is necessary to define and determine workflows in relation to the milestones. For the fulfilment of the relevant objectives, sanctions should be laid down in the event of non-compliance with these agreements. B) The level of interaction with each other - this is where the way in which the DataPLANT participants wish to interact and work with each other and how conflicts should be dealt with should be recorded. C) The level of the participants responsible for management - this will define the principles and rules, for example, the spokesperson should follow in their dealings with the members of the association.

The *providers* need a sound financial base to offer sustainable service and keep long-term promises. They agree to be open to review processes and user feedback and have to transparently report on the expenses, i.e. costs should be reasonably justified. In addition, they rely on (fully/partially) funded grace periods for changes in service structure to allow the fade out of deprecated services and the ramp-up of novel ones. Ultimately, they report to the senior management board and the general assembly. Non-compliance will result in reduced disbursed funds. The *data champions* and *users* agree to principles of Open Science, Open Data and good scientific practice of the DataPLANT community. They will have regular access to data stewards, education programmes, data repositories, handles for data citation and will get rewarded for exemplarily annotated data sets (overseen by the data stewards board). Funds for *developers* hosted at the applicant institutions or participants are not passed on in advance entirely. 20 percent will be held back until the completion of agreed-upon milestones. The progress and setbacks will be reported regularly to the senior management board and on a yearly basis to the general assembly.

Project particularities. DataPLANT faces a number of challenges that need to be addressed through project management and governance. The persons involved in DataPLANT - both the researchers explicitly hired for the project as well as the research groups involved - are subordinate to their respective institutions, which means that personnel responsibility is

distributed. The participants come from different institutions and are therefore structurally autonomous, at least in part. Participation in DataPLANT is therefore not necessarily of equal importance for all participants, as they are mainly dedicated to their ongoing research projects. Nevertheless, DataPLANT must ensure that joint results are achieved, for example in standardization or ontologies, and that there is close cooperation between the research groups even if they are in competition with each other for findings and third-party funding. The governance of DataPLANT must ensure that decisions are made in a timely manner, despite perhaps multi-tiered and decentralised structures. Expected discussions must be moderated by the governance structures. The participants in DataPLANT have strong self-interest, which is accepted by all participants. The NFDI can only exert as much influence on individual projects as is permitted or agreed in advance (e.g. for the provision of data and the willingness to participate in the committees).

2.6 Operating model

The preliminary DataPLANT operating model, which will be evaluated and refined during the project runtime, rests by and large on three pillars: A) The NFDI funding of personnel hired for support, development and operation, B) the in-kind contribution of hardware and services by the applicant institutions and additional funding, and C) in-kind contribution of personnel.

Support, development and operation. The funding of DataPLANT compensates for on-site data management support, development, coordination and organisation costs of the consortium. To meet the expressed needs of the community for support in data management and supportive services, significant funds will be set aside for data stewards. We plan to provide about 10,000 person hours per year, in a flexible and fair on-demand model. Organisationally, the data stewards will be hosted at different sites all over the country, attached to larger research groups but independent of these, while being paid through the applicant institutions (UFR). They are dispatched from these locations to all other participants and the wider community. Another significant proportion will go into developers and personnel to support and moderate standardization, metadata definition and development of the DataPLANT Hub as a central entry point to all relevant research data management and workflow services. The proposed governance structure ensures the alignment of the planned work programme with the users' expectations. The **in-kind contribution of hardware and services** stems from the resources provided by the hosting institutions in the form of qualified research data management personnel, supporting administrative personnel and local infrastructure. Further contributions are the BinAC HPC cluster in Tübingen, the bwCloud infrastructure in Freiburg, and the upcoming BW Storage-for-Science services both in Tübingen in Freiburg significantly co-funded by the state of Baden-Württemberg,

the DFG and the hosting institutions. The de.NBI partitions in Freiburg and Tübingen - compute and storage resources shared by different bioinformatics communities. - are financed by the BMBF. The **in-kind contribution of personnel** brought in by the DataPLANT partners in the form of the co-speakers as well as scientific and supporting administrative personnel both for the hardware as well as for project administration.

Service and infrastructure operation model. DataPLANT has a layered understanding of services: Base-level services mainly of compute and storage are operated and contributed by the service providers. The to-be-created higher level services help to close the gap between users and providers. The gap consists both of services tailored to the particular needs and necessary training and qualification to make the most of existing and novel services in the field of plant research. Significant efforts which are not covered by the provider's service stack and the researchers' project budgets will be spent in DataPLANT in this domain. They will be linked to incentives fostering the objectives of the consortium. For the infrastructure heavy compute, storage and higher-level workflow services there will be a mixture of funding streams used. The access to DataPLANT services follows the model of free basic services to offer low hurdles for access to every member of the plant research scientific community. These basic services are provided out of DataPLANT funding and the in-kind contributions of the partners. To avoid the overloading of existing resources requirements beyond the basic level would induce forms of resource transfers to compensate for the extra efforts. A major challenge in the operation model will be the compensation of efforts as only in an ideal and static environment would the resources brought in perfectly equalize with the efforts spent by partners and providers.

Compensation models. We acknowledge the fact of challenges regarding cost compensation models which comply with non-profit/public-benefit requirements and capable of being integrated into future NFDI governance structures. Calculation of costs and refinancing becomes unavoidable at some point to allow sustainable cooperation. Different funding streams need to be taken into account. Every institution has funds to pay for commercial third-party services and consulting, but it is nearly impossible to proper receive such funds as a research institution.

Direct flows of money in a consortium of differently organized and funded research institutions is an issue to be iterated and solved via the corresponding cross-cutting topic. The data steward services can be clearly accounted for. Thus, it could be an option to use vouchers or coupons on services in exchange for redirected financing via a grant application. An endorsement model could be established to foster such developments, where research funding agencies see the NFDI as a service broker. The NFDI structure ensures good scientific practice by offering certified services which are in turn applied for by research groups. The funding agencies would divert a certain

amount of support to the NFDI in relation to the equivalent of the services requested. A base level funding e.g. either through universities or the NFDI would compensate for consultation to non-successful applications. Such options need to be discussed and developed together with all stakeholders within DataPLANT, the general NFDI level, and the appropriate political sphere. Efficient ways of reimbursement without overhead and bureaucracy are needed to offer a sustainable model for long-term cooperation and viability. The outlined concept including the NFDI as a registered society could be a feasible way for a non-profit oriented operation model.

3 Research Data Management Strategy

Research data management is a multi-faceted endeavour involving a wide range of stakeholders. Primarily it is meant to improve the good scientific practice of researchers and generate high quality and reproducible results. Every researcher is expected to be aware and follow the standards set by their scientific domain, agreed upon by their research institution and required in scholarly communication. While various consortia have made suggestions on best practices and processes towards fulfilling these principles, it is nevertheless always up to individual researchers' initiative to adhere to them. As a result, comprehensive information of the required quality for use by third parties is only available in exceptional, rare cases.

Modern plant biology involves the integration of multiple heterogeneous data sets across all system level, in order to understand the underlying physiological responses as a highly interconnected molecular adjustment [34–36]. Therefore, complex experiments using different technologies, such as proteomics, transcriptomics, and metabolomics, are necessary by default and generate data of various types (quantitative, qualitative, text, computed values) in diverse formats and in ever-increasing abundance. As in other scientific domains, plant science has become an interdisciplinary research area to collaboratively investigate pressing research questions of tremendous importance. Crop production, food security, climate change resilience, healthy nutrition, and sustainable agriculture build upon the understanding of the underlying molecular mechanisms of the plant system. Even for basic interpretation, modern high-throughput technologies used in plant science require computationally expensive data processing and a combination of measurement and reference data. For example, during processing of proteomics data acquired by mass spectrometry, recorded spectra are commonly evaluated against a known set of known protein sequences expected to be present in the respective organism according to a reference gene model comprising basic functional annotation. Here, the scenario becomes evidently challenging for the individual researcher. Considerable knowledge in multiple research fields is required to compile complete metadata for a meaningful description of the experiment. Additionally, a study will often include multiple measurements or assays each of which needs to be integrated with reference data, requires various processing and computational steps and

dedicated data publishing procedures. This leads to the central point of our requirement analysis resulting from our DataPLANT survey: Researcher need practical support to cope with the fragmented and overwhelming landscape of information sources to enable the democratization of research data. In the following we will discuss currently existing and overlapping information infrastructures, data repositories or reusable software sources we will homogenized, integrate, provide 'best practices', and build upon in DataPLANT.

In plant research there are various information resources and data portals of very high quality. UniProt[37] and Ensembl plants[38] are integrative resources presenting genome-scale information for a growing number of sequenced plant species. Additionally, PLAZA[39] provides an integrative resource for functional, evolutionary and comparative plant genomics. Data portals and specific databases like the The Arabidopsis Information Resource (TAIR)[40], Araport[41], Aramemnon[42] provide fine-grained species-specific reference knowledge. However, all the above-mentioned resources have a certain overlap regarding the information they provide and it is not evident to a user how to cope with inconsistent information when queering multiple resources. This is especially true when we also include knowledge bases on gene product functions, localization and association into the picture. Besides the challenges of information retrieval resulting from the heterogeneous landscape, researchers are often resilient to participate in the update and release process. Currently, manual curation is the classical procedure to integrate individual research results into reference knowledge platforms as the individual research is not directly connected to the information provider. A possibility to connect individual research results directly with data portals is through well annotated research objects. Starting at the experiment level, there are different open lab book implementations available to replace traditional lab books with a digital and shareable version. However, only in combination with the usage of Research Object[43], Research Object Crate[44] or ISA data model[45], electronic lab notebooks provide the rich description of the experimental metadata (i.e. sample characteristics, technology and measurement types, sample-to-data relationships) that make the resulting data and discoveries reproducible and reusable[45].

Best practice for the core data suggests the selection of a technology-specific data repository. ProteomeXchange[46], Gene Expression Omnibus (GEO)[47], SRA/ENA[48] and Metabolights[49] are well established data exchange platforms that enforce metadata annotation tailored to the individual technology. In contrast, generic data repositories like figshare[50] and Dataverse[51] do not require a technology-specific and laborious annotation process, but in turn do not ensure the necessary metadata annotation. For correct interpretation, and thus replicability, comparability and interoperability, a user needs to make sure that minimum information requirements are met. In the plant field, excellent standardizations for experimental data collections are the 'Minimal

Information on Biological and Biomedical Investigations'[52], 'Minimal Information about a Plant Microarray Experiment'[53], 'Minimal Information about Plant Phenotyping Experiments'[54].

Beyond these, a considerable further amount of knowledge about various ontologies, thesauruses, standard definition, compute and storage option is necessary to provide adequate metadata annotation to the research data[55]. However, it is extremely challenging for an individual researcher to acquire the relevant skills for using or curating repositories, but also to allocate the resources and capacity to actually do so in daily research practice. In addition, many researchers view data as sensitive research output that could easily be misused or mis-interpreted when taken out of context. Thus, many scientists do not trust global repositories unless they have direct and personal connections to these researchers' own work or find it too time consuming to validate their trustworthiness. Overall, the landscape of research data management infrastructure – not only encompassing data storage but also computation – appears to researchers as fragmented and overwhelming. Simultaneously, technological progress has increasingly rendered research in plant science a team effort that makes a dialogue between data producers and consumers more and more important. Thus, standardized, easy-to-use means to enable exchange and reuse of research data are of utter significance.

DataPLANT envisions a solution that allows research data management with minimal additional effort and a system to foster the incentive of the researcher in the plant community.

Envisaged state of research data management within the scope of the consortium. A central insight underlying the research data management (RDM) envisaged within DataPLANT is that due to the quickly changing nature of research methodologies and workflows – especially where computational methods are concerned – core aspects of the proposed RDM strategy must be adaptive and constantly evolving. The RDM strategy of DataPLANT is therefore aimed at facilitating the community-driven continuous evolution of standards and processes. At a high level, DataPLANT's RDM strategy consists of the following components:

- A centrally coordinated, community-focused, and requirements-driven **standardization process** for metadata and research workflow annotation, to derive a standard that serves as a basis for all annotation, quality, and storage efforts within DataPLANT. The developed standards will leverage internationally recognized meta-standards such as Research Objects[44], the ISA model[45], and the plant biology-specific MI* ontologies.

- A centrally coordinated, community-focused, and requirements-driven set of easy-to-adopt **(best) practices** for effective adoption of the proposed standards and services into research practice. Most importantly, these processes will be designed to ensure

adherence to the FAIR principles, allow for substantial automation in quality control and curation, and encompass a set of incentives towards their adoption.

- Personnel in the form of a centrally organized pool of **data stewards** to provide on-demand assistance and counselling on research data annotation and processes, while ensuring continuous documentation of requirements and problems. Supplementing the DataPLANT standards and process model by dedicated personnel is in our vision a central requirement for effective adoption of DataPLANT's RDM strategy and an effective channel for engaging broadly with the user community.

- Software services available through the **DataPLANT Hub**, a science gateway to engage with the user community, facilitate adoption of DataPLANT practices with minimal overhead, while ensuring systematic monitoring of standards and practices as well as their adoption. In brief, the hub will provide the software services needed for effectively working with the envisioned practices and standards, and act as a data repository for DataPLANT.

- Base infrastructure in the form of **storage and compute infrastructure resources** that support the higher-level services of DataPLANT and ensure its long-term viability without external dependencies.

- Services to support sustainable **FAIR long-term access to complete research contexts**. They complement the FAIR efforts on data sets by filling a gap between base level bit preservation and technological advances in software and hardware.

- A comprehensive **training and education program** for the dissemination of DataPLANTs RDM strategy and general RDM best practices, tailored to the requirements of specific interest groups (students, researchers, project managers).

We describe these service components in detail in chapter 3.3.

Monitoring user needs and evolution of DataPLANT strategy. The DNA of DataPLANT is user centric; the organisational and governance model includes several platforms for participation and involvement. The designated scientific community has a strong representation in relation to the provider and developer core. It is designed as a learning organisation to allow for organisational and structural changes and enhancements. Regular meetings of the general assembly and the boards as well as the creation of special interest groups (SIG) allow the permanent evolvement of DataPLANT [Figure 4]. As a direct and regular link between the users, providers and developers specialized data stewards are operating. They interact with the scientists and research groups in the field and monitor user requirements and needs. They accompany the

evolution of ontologies in use and the evolvement of standards. Through specific working groups and the DataPLANT office they interact with the international plant research and wider bioinformatics community. As a permanent body the scientific board maintains the scientific oversight and provides feedback on technical and scientific developments to the general assembly as well as the office for further coordination. The general assembly ratifies all strategic decisions. The senior management board provides administrative oversight, discusses, evaluates and develops business and organisational models to manage the project and the various resources. The management board is in contact with the general NFDI, the other consortia for the cross-cutting topics and the relevant stakeholders nationally and internationally supported by the office. It coordinates the change processes.

Data selection and quality management. Active participation and user involvement represent the essence of DataPLANT. Therefore, we are planning to accept data of different quality. When accessing data quality in the context of RDM, we will differentiate between measurement quality and data completeness. Due to design principles, data processed, analysed and managed using the DataPLANT service infrastructure, our best practices ensure to achieve metadata completeness during the process. Additionally, the scientific workflow platform Galaxy will be used to include quality control checks after every step in a given workflow, ensuring a transparent report of the measurement quality but also potential analysis errors as early as possible. A unified reporting system will provide comprehensive quality measurements, including the entire provenance of the analysis. At the stage of journal publication or public referencing the data object, metadata standards and completeness can be automatically enforced using minimal standards. In DataPLANT, data stewards can closely monitor repository usage, and drive quality improvement at the user sites. Further, we envision data quality and metadata completeness as an incentive to get resources (storage, access to data stewards) from the DataPLANT service infrastructure**.**

3.1 Metadata standards

In the past years, many standards have been developed for describing metadata across a variety of research domains. Some of these standards are very generic and domain-agnostic (such as e.g. the metadata subset of the EU's CERIF – Common European Research Information Format), while others specify metadata for particular domains (e.g. the EML – Ecological Metadata Language). Most metadata standards address annotation needs at three levels: conceptual, logical, and physical.

Among the many existing standards, the international standard Research Object Crate[44] (RO) is of particular interest for DataPLANT. ROs are aggregations of resources with rich semantics that

allow the succinct description of data, methods, and people in a scientific research context. Based on widely used standards and basic ontologies such as schema.org and RDF, ROs are designed to be extensible through supporting domain-specific and use-case specific ontologies. A core concept of the RO standard are profiles that combine a representation of metadata with an expected set of resources and specific vocabularies (or further specifications) to be used for annotation. In other words, profiles indicate the purpose of a particular RO and specify assumptions to be relied upon in comprehending an RO. Furthermore, the RO specification describes various approaches to the packaging to RO for archival, such as the Research Object Bundle[43] that represents a collection of ROs, or the RO crate that uses a more widely interoperable format. While providing a rich framework for research metadata and its archival, the genericity of ROs makes its application in a particular domain challenging without further standardization efforts. Two often-encountered difficulties in the application of metadata specifications are exemplified by ROs: i) While ROs are very general, it is exactly this generality and a lack of specific vocabularies complicating their application to specific domains in practical experience, and ii) representations are aimed at machine readability and are thus, without proper support by software systems, exceedingly complex to manage for (typically non-expert) researchers.

A more process-focused metadata model is described by the ISA[45] (Investigation, Study, Assay) standard that is used in biological science more often. The ISA standard itself describes a data model as well as serializations (e.g. tabular or JSON), and is accompanied by a rich, open source set of tools that allow researchers to work with ISA metadata, thereby addressing the usability gap encountered e.g. by ROs. The ISA model in itself does not specify a sufficient set of workflow attributes that would allow it to comprehensively address metadata cultivation in the context of computational biology workflows, nor does it provide a vocabulary specific to plant biology.

The above-mentioned metadata standards and schemas will be compared and matched against PREMIS, a high-level Metadata standard for digital preservation. A closer look will be taken on technical Metadata, Provenance Metadata, and Rights. Technical Metadata should give proper information about the data itself, the level of preservation and preservability, the inhibitors for reusing data. Provenance Metadata describe the actions taken from agents and software on data; they document the history of data and to guarantee the chain of custody of digital objects. Rights Metadata describe the copyright status or moral rights, the terms and conditions by law or contract and the licenses under which data may be used or distributed.

Focusing on the computational processing of data, metadata can also be utilized to annotate corresponding workflows, with the ultimate goal of documenting not only annotating research

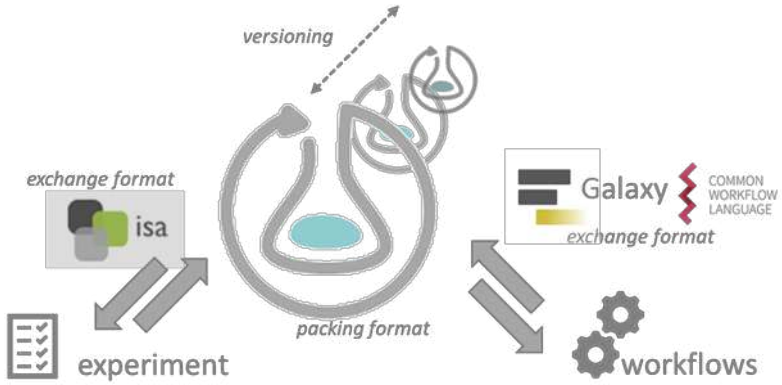

Figure 5 Overview of the exchange model in DataPLANT. Data are imported and exported using the most user-friendly ways and stored and versioned in an appropriate packing format (RO). This ensures compatibility and interoperability with international standards and other research areas.

data, but also its evolution towards scientific insight in a manner that allows accurate reproduction, indexing, and execution of such workflows and furthermore documents the overall provenance of scientific results. For example, the Common Workflow Language (CWL)[56] is a standard for capturing and describing analysis workflows and tools in a way that makes them portable and scalable across a variety of computational environments. CWL is aimed at supporting data-intensive science. A key difference between annotating data and workflows is that annotation for the latter can be largely automated, reducing metadata quality problems in general. However, for this to function effectively, domain-specific peculiarities such as the use of specific analysis codes or customary processing steps must be accurately and comprehensively describable. This necessitates a corresponding vocabulary that is not yet available for the domain of plant research.

For correct interpretation, and thus replicability, comparability and interoperability, RDM needs to make sure that minimum information requirements are met. In the plant field, excellent standardizations for experimental data collections are the 'Minimal Information on Biological and Biomedical Investigations'[52], and 'Minimal Information about Plant Phenotyping Experiments'[54]. However, we get a grasp of the complex landscape when we consider the minimum information requirements for assays used in plant biology like MIAME[57] (microarrays), MIGS[58] (genomics/metagenomics), MIAMET[59] (metabolomics) and MIAPE[60] (proteins) with its sub-standards: MIAPE-MS (Mass Spectrometry), MIAPE-MSI (Mass Spectrometry Informatics), MIAPE-Quant (Mass Spectrometry Quantification), MIAPE-GE (Gel Electrophoresis), MIAPE-GI (Gel Informatics), MIAPE-CC (Column Chromatography), MIAPE-CE (Capillary Electrophoresis), or MIMIx (Molecular Interactions).

In DataPLANT we will build on the existing standards to preserve compatibility and interoperability towards the international communities and maintainer of these standards. However, we envision an operational fusion and homogenization of existing standards to benefit from individual advantages but increase the usability by decreasing complexity and effort on the user site. We will use RO as a basis storage point of aggregation, ISA-TAB for user interaction as researchers

are familiar with the use of spreadsheets, and CWL61 and Galaxy as the workflow descriptor [*Figure 5*]. For usability we will homogenize the Minimal Information landscape and base RO profiles on the existing efforts. We will customize the until now generic metadata standards to the needs of plant research community and invest in an optimal vocabulary compendium without sacrificing compatibility to international standards. DataPLANT will take part in the international standard and vocabulary development efforts by reporting changes and connecting to the user. The DataPLANT consortium is also engage in metadata standard communication across all NFDIs as part of cross-cutting topics. Moreover, annotation of metadata itself is of central interest when developing metadata cultivation workflows (e.g. enrichment, quality improvement). In this context, metadata itself must be the target of annotation, for example to express versioning and changes. A corresponding vocabulary is within the scope of DataPLANT's metadata standardization efforts.

3.2 Implementation of the FAIR principles and data quality assurance

DataPLANT follows a clear strategy for data use, access, findability and reusability in accordance with the FAIR principles[1] by embedding it on all relevant layers of its research data management strategy. Its approach delivers FAIR compliance by design: A core database embedded to the DataPLANT Hub ensures persistent and unique identifiers for all entries covering the research context as well[29]. DataPLANT is building a layer on top of trusted infrastructure and repositories that themselves ensure accessibility. We develop a dedicated version control semantics based on current data standards and formats, ontologies and information requirements to ensure interoperability. Due to the proposed tracking and linking strategy, we gather accurate information on provenance. We will advance the concept of the complete research context and develop a sustainable service for accessing data in past operation contexts (of deprecated software and hardware). Further, DataPLANT will enable exporting snapshots of the project using Research Object as an international exchange format for guarantied reusability.

1. For data sets to be *Findable* the DataPLANT services will offer advanced search functionality and advertise its datasets to data aggregators e.g. OpenAire[61]. The DataPLANT HUB provides a searchable resource and access to the data publication service. The service will provide necessary IDs and require to add rich metadata through specification.

2. For data sets and workflows to be *Accessible* the DataPLANT Hub provides a platform which links to all relevant resources and by allowing access to data and metadata through unique identifiers and stanard machine-readable protocols. The consortium further advocates for Open Science and Open Data.

3. DataPLANT will support the *Interoperable* principle by starting from the international community established ontologies and metadata sets. It will be further ensured by the focus on Research Objects and the work package dedicated for standardization. One of the objectives of the consortium is that data is well enough described to be potentially useful in other disciplines. As this cannot be known in every aspect it will be negotiated through cross-cutting activities and thus be an iterative process.

4. For data sets and metadata to be *Reusable* will be ensured through the specification developed from domain-relevant community standards. The research context and provenience is kept and can be reproduced at any time. Further will the DataPLANT Hub allow for different ways of access e.g. for further reuse of data through machine learning, streaming or further use cases.

The DataPLANT approach ensures further structured organisation across different standards, requirements and types as well as autonomies resources and allows data structuring over time and multiple collaborators. Thereby, successful collaborative work will encourage adequate data annotation with minimal effort.

Most of today's services focus is on access and sharing (collaboration) while long-term accessibility is usually ensured through bit-preservation procedures and some format-specific file format migration services. If FAIR is the success criteria for successful long-term access and reuse of data sets, a broader and technically more diverse approach is required. Our research data service cannot simply refuse badly rated formats (or data sets containing such files). Such highlighted risk should be the starting point for a productive approach. For this the data creator should be involved and the potential access and reuse issues of his data set should be discussed and step-by-step improved. The file versioning service offered within the DataPLANT services stack is a basis for such an endeavour. A dedicated service acknowledges the fact that (meta)data quality will evolve over time with community interaction.

The services stack in DataPLANT will provide the necessary interfaces to other services (within the NFDI) to support interoperability. It will communicate with registry services of research data repositories like R3DATA. Further on, DataPLANT will put attention on sustainable base level services for storage and compute which are required for long-term stable access to well referenced and published data sets. From the very beginning of a research project, a data management plan should provide information about discipline specific metadata and related vocabulary needed for the enrichment of data objects. Especially the continuation of a provenance information chain on data objects throughout the life cycle is a key aspect of metadata management.

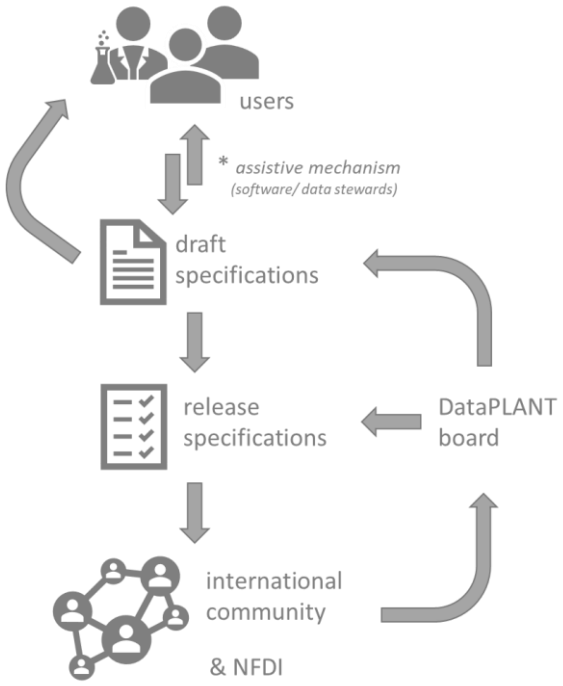

Figure 6 Overview of the DataPLANT specification management cycle. The approach ensures a direct usability of draft specification by the users and facilitates the interaction.

DataPLANT tackles quality assurance by multiple measures. Several types of quality metrics are taken into account: Metadata completeness and quality, measurement and data quality, the use of open data formats including the use open source base processing tools and plant research specific quality measures. We will provide easy to use tools to automate high-quality data acquisition. The design and implementation QA processes for FAIR foster data quality. Finally, data sets will be automatically FAIR when following the agreed-upon procedures. Through community engagement we will ensure that processes will be researcher friendly. The verification process for data sets combined with the versioning approach and the workflow verification methods getting implemented for Galaxy allow for a broad data and workflow quality assurance. The technical means are complemented by the assistance of researchers through the data stewards and the evaluation of services provided through the DataPLANT feedback mechanisms [*Figure 6*]. These activities are overseen by the scientific and technical boards to ensure a structured evolvement of the implementation of the FAIR principles and the quality assurance of research data. Further on we will implement an internal scoring system, incentive structures, and rewards within the consortium.

The verification service for data sets combined with the versioning approach plus the workflow verification methods getting implemented for Galaxy allow for a broad data and workflow quality

assurance. The technical means are complemented by the assistance of researchers through the data stewards and the evaluation of services provided through the DataPLANT feedback mechanisms. These activities are overseen by the scientific and technical boards to ensure a structured evolvement of the implementation of the FAIR principles and the quality assurance of research data.

3.3 Services provided by the consortium

DataPLANT will offer three types of services: Moderation of various standardisation practices, direct support services provided through the dispatching of data stewards and various infrastructural services offered through the DataPLANT Hub and the providers. This encompasses centrally coordinated, community-focused, and requirements-driven set of easy-to-adopt practices for effective adoption of the proposed standards and services into research practice. Most importantly, these services will be designed to ensure adherence to the FAIR principles, allow for substantial automation in quality control and curation, and encompass a set of incentives towards their adoption.

DataPLANT metadata standardization and RDM practices. DataPLANT's mission is to develop comprehensive metadata standards that address the specific needs of metadata cultivation in the plant biology research domain. While it would be possible to simply leverage an existing generic standard for this purpose, these are lacking specificity. The general nature of annotations available through general standards captures only a fraction of the annotations that are needed to ensure full reproducibility and interoperability of plant research data. However, we will leverage the large body of effort that has been expended on behalf of generic specifications by basing our specifications on the existing Research Objects standard. It is sufficiently generic to support the vocabulary and representations to annotate plant biology research data, workflows, and metadata; on the other hand, it is specified in sufficient detail to allow immediate operative use. With its profile mechanism offers an effective instrument on which we will base further uses of our standard, e.g. towards metadata cultivation and quality control. In this sense, DataPLANT will channel existing but fragmented efforts underway in the plant research community to drive the evolution of a common standard. This DataPLANT metadata standard will consist of three interlinking specifications:

Data annotation: A specification for the annotation of research data sets with metadata. Based on the ISA model, we will create a specific vocabulary within the Research Objects framework and a corresponding profile that captures all domain-specific annotations required in plant biology.

Workflow annotations: A specification for capturing experimental and computational workflows and data provenance specific to plant biology. Based on the CWL provenance profile for Research Objects, we will develop a specific vocabulary and profile for annotating plant biology computational workflows.

Annotation of Metadata: A specification for the annotation of metadata itself and its evolution. Again, based on the Research Object framework, we will create vocabulary and profile to annotate metadata evolution, specific to the use cases of plant research, and allowing automation to a large degree. To ensure community-wide acceptance and relevance, these standards will be initially developed following a stringent requirement engineering process involving the DataPLANT user community and evolved through feedback gathered by the data stewards. As data and metadata curated among DataPLANT partners must be interoperable with other internationally accepted standards, such interoperability will be considered as a primary component in the creation of the DataPLANT metadata standard. Here, we can leverage existing interoperability specifications for Research Objects, but will furthermore consider further interoperability mechanisms with e.g. CERIF and other widely used standards. The standards developed within DataPLANT will be fully open, published, and will be contributed to the international community as a Research Objects substandard. This process has been successfully conducted in other communities, e.g. for regulatory science (BioCompute) or Digital Preservation (BagIt). We will furthermore institute processes and collaboration models for the international plant research community to engage with, utilize, and strengthen DataPLANT's standards. While the DataPLANT standards will initially be developed and adopted among DataPLANT's user community, our goal is to develop standards that benefit the international community of plant researchers in the long-term. In coordination with other NFDIs (declaration of Berlin) we will contribute to the development of NFDI-wide (core) metadata standards and implement it into our standards. We envision to actively contribute to and leverage from the efforts of other NFDI consortia that develop similar standards in other domains, as common problems, approaches, and solutions are to be expected in adapting typically generic standards to specific domains or formulating these anew. Here, we will actively participate in efforts that seek to address these common challenges cutting across all NFDI consortia and will ensure that solutions found at this level will be reflected in our standards.

Data stewards. We envision data stewards as a core element of DataPLANT RDM strategy. They support the community directly in their daily data management tasks on a regular base. The research groups directly profit from the support. Thus, they play a special hinge role between service providers, individual researchers, groups and the wider community. By providing support and advice on data and workflow management, they foster standardization both of metadata,

ontologies and data handling. They close the gap between single researchers, research groups and the wider community as well as the gap to the technical systems. The coordinated use of data stewards supports the adoption of good scientific practice among the addressed community. Through the regular direct interaction with the users, data stewards can provide rapid feedback to the service providers, the DataPLANT board and the wider community on needs and requirements.

Each data steward provides roughly 1600 person hours per year. 200 for self-qualification, participation in RDM and community specific conferences and workshops. 200 for coordination within the consortium, user support, reporting back from assignments, regular scheduled meetings e.g. among the data stewards. There will be 8 times 1200 hours in the data steward mission pool per year after an initial ramp up phase. These roughly 10.000 hours will be distributed as follows: For initial startup each group or individual gets data steward time equivalent to the amount of data. A mini application - giving information on data, amount, type, research project goals - is required through a web form. The median expected initial support is 100 hours per incident. In the first assignment, data steward will help to implement the relevant workflows into the participant labs according to data types, training on data management and metadata best practices following the agreed upon NFDI standards. To ensure quality standards and fair distribution of support evaluation criteria for data steward requests will be applied: Initially (first call) low hurdles are applied; inquirers just have to give some preliminary information so that the stewards can be chosen by expertise and preparation of the initial assignment and courses or workshops can be prepared to train the requesting group. The request evaluation primarily lies with the group of data stewards. Special requests, conflicts which are not solvable on that layer will be passed on to the Senior Management Board to decide. Additionally, this body takes steering responsibilities and may adapt the distribution if necessary, after a ramp up period followed by an evaluation of the process. We assume a decreasing demand from single participants but expect rising request from the wider community. In further phases the delivered, annotated and published data sets entitle participants for additional allowance for further data steward support. To ensure productivity, the applications will be evaluated according to an agreed upon distribution and resource allocation algorithm. Additionally, requests can be supported through co-financing by the participants or new members or own personnel occupying similar workplace descriptions can participate in the data steward team and board.

The DataPLANT Hub is the central science gateway supporting the research data management practices developed within DataPLANT and automates them wherever possible. Its realization will consist of a web service for working with research contexts (linking to data, and directly containing or linking to external metadata and workflows as appropriate), with a corresponding backend that is connected to the DataPLANT infrastructure resources. Centrally, it will initially provide the following services for research contexts:

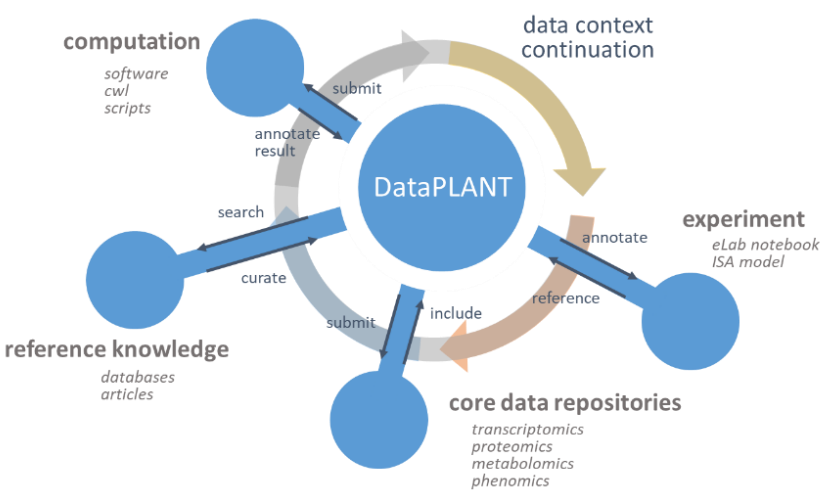

Figure 7 The DataPLANT Hub accompanies and supports the complete research cycle. Data context continuation manifests that experiments will influence knowledge and following experiments.

- Automated annotation of data and workflows within research contexts based on the standardized ontologies developed in TA1
- Searchable access to data, metadata and workflows
- Access to a recommendation system for metadata content
- Versioning and provenance, with automated provenance tracking where possible
- Automated quality assessment including data and metadata quality, and communication of quality to users (“data traffic light”)
- A mechanism for publishing specific versions of research contexts with guaranteed FAIR compliance
- Mechanisms for collaboration, such as sharing research contexts with other researchers
- Facilities for converting metadata to and from DataPLANT’s formats and vocabularies to those of other repositories of relevance to the user communities

- Mechanisms for dispatching computational workflows on the data contained within research contexts, both on the computational resources provided by DataPLANT but also on those provided by third parties

In brief, the DataPLANT Hub [*Figure 7*] will serve as a central platform for realizing all relevant RDM practices. There will be a direct connection to analyse plant research related data, embedded into the European Galaxy server (ELIXIR/EOSC) and running on the bwCloud, the de.NBI Cloud and further resources. We are aiming at providing and maintaining over 200 relevant tools for our community with access to more than 100 plant reference genomes and transcriptomes.

The Storage and Compute infrastructure resources provide the basis for the operation of various research data management services through the DataPLANT Hub as well as offering the necessary capacities to run training and education sessions with all necessary modules. As DataPLANT strives for Open Data and Open Science endeavors, it has to provide a solid base level infrastructure. Such a common infrastructure allows an easier sharing of data sets and novel IT-based insights envisioned by the NFDI initiative. Additionally, the infrastructure serves as an attractor to well annotated data sets and provides an incentive to make data sets FAIR. A common infrastructure creates a focal point to federate local resources. DataPLANT storage resources and repository services provide a viable alternative to the lock-in business models of many publishing houses. Not all plant data resources available nationally and internationally seem reliably sustainable, thus it has to be ensured that at least a local copy of relevant data is kept. Every service and online resource has its limits regarding compute or storage capacities. DataPLANT will alleviate that problem by providing significant own resources brought in by the participating institutions combined with existing resources from initiatives like de.NBI and EOSC.

Various higher-level services like repositories or data versioning will be offered in a federated way by the participating IT centers in DataPLANT. The storage will be provided in different ways ranging from filesystem services to object storage in different levels of redundancy through the Baden-Württemberg Storage-for-Science (bwSFS). bwSFS is presently under construction at the Universities of Tübingen and Freiburg and will become available mid-2020. This is a general-purpose research data management infrastructure under the responsibility of the local Research Data Management Groups with strong ties into the Bioinformatics communities. The system will allow, among other things, to specify lifetimes and importance of data sets. Data marked as important will be backuped georedundantly between different sites. The long-term archival partition of the system will guarantee availability of data set for lifetimes above 10 years.

The compute services - offered as cloud or HPC infrastructure - will provide the necessary resources for research workflows and the generic data handling required e.g. for visualization, aggregation, metadata annotation, sophisticated searches, indexing and quality assurance. Resources like the BinAC HPC cluster in Tübingen and the de.NBI clouds in Tübingen and Freiburg were designed to suit the needs and requirements of the bioinformatics community. Both service providers implemented modern hardware deployment schemes for their HPC and cloud infrastructures which allows a dynamic allocation of resources. Containerization and virtualization allow for differentiated and separated software environments which cannot interfere each other allowing research groups independent setups and adaptations to their particular needs. Special purpose hardware like GPU accelerators can be flexibly assigned. The already enabled authentication and authorizing infrastructure allows non-local users access. Galaxy allows for a flexible resource scheduling independent of a specific location. More advanced scheduling can consider various parameters for compute job distribution including data and machine location as well as the user affiliation.

FAIR long-term access to research contexts. DataPLANT follows a holistic approach for preserving the complete research context by providing necessary services, interacting with the international digital preservation experts and offering these services to other scientific communities within the NFDI. The concept of research environment or research context encompasses the software stack, explicit description of workflows, custom scripts, and settings a researcher used for processing the data sets, going beyond a set of descriptive metadata for a given data set. As these components of the research environment are stored on computational resources, they should be also considered as data. As a major consequence of this, the whole data life cycle, as well as the FAIR principles, are applicable to the research environment. Leveraging the data life cycle, research contexts should be versioned. This means, it should be noted which version of a research environment was used to create a specific result. Also, research environments should be archived, in order to enable reproducible computations. DataPLANT will advance these important aspects of reproducible research environments and suggests methods for combining them with an organized data and metadata management.

The explicit notation of an analytical workflow in a workflow language is good scientific practice. Publication of these workflows according to FAIR principles is essential to ensure transparency and reusability[1]. Analogues to Open Data, open research environments increase reproducibility and - favorable for an author - also increases the number of citations[62]. But a well-defined workflow is just a starting point for achieving reproducibility of a computational pipeline[15,63]. It is obvious that a simple listing of software used by this pipeline is not enough for other researchers to reproduce a pipeline[64]. Software containerization techniques introduced in recent years like

Docker and Singularity[65] allow researchers to package a specific software stack into a virtual research environment[66–68], including an operating system and additional data like custom scripts, in a single entity called container image. This technique enables versioning and archiving of research environments and pipelines, as the environment is bundled in one image. Bundling the research environment together as software container increases the reproducibility of computations using the bundled methods.

Our previous findings support the point that some extra attention to the data's software dependencies is necessary[69]. Different files and file formats may have software interdependencies concerning reuse and thus, long-term access planning should take these into account. Therefore, DataPLANT services will implement measures for publication and sharing of scientific workflows to improve and support reproducible research practices. It also enables researchers to share this single entity via scientific data repositories, public container hubs, or institutional repositories residing on systems such as the federated bwSFS. On the conceptual level the planned standardization will foster the findability and through better compatibility the reuse of findings. This will be supported by the service infrastructure in the form of the DataPLANT Hub to allow for advanced search and indexing. As a crucial building block of the DataPLANT services and a contribution to cross-cutting activities we will further improve and operate a long-term access service to (outdated) research environments. It will include software preservation coordination in the national and international context, in particular we plan to cooperate with the Software Heritage Foundation in order to maintain access to source code and source code repositories and extend our ongoing cooperation with the Software Preservation Network to ensure access to critical software components. To be able to assess the long-term access and reuse risks of data sets, we will extend upon an implemented and deployed data set characterization prototype, using a simple traffic light visualization, signaling the user the long-term reuse probability of a given data set and file format. The results of the characterization service can be used either as pre-ingest check, e.g. as a tool for feedback to an initial submission, i.e. flagging unsustainable, unknown or otherwise difficult file formats. Based on this feedback, individual researchers will be advised by the data stewards to reconsider their file format choices (if possible) and their awareness can be raised on the un-sustainability of their format choices. These procedures use the file versioning service to allow a gradual improvement and public feedback. Furthermore, the characterization results will be used to guide a software collection, required to render certain data sets and as an input for the research context preservation strategy.

Training and education program. To support the development of good scientific practice in the field of plant research and to involve all relevant groups from project managers and principal investigators to PhDs and students a comprehensive training and qualification program is

required. We will provide and update in coordination with other consortia and the general NFDI suitable training infrastructures and (online) training materials. The training and education program includes indirect measures like the extension of the relevant curricula on research data management aspects as well as workshops and summer schools. Qualification on data management plans provides the necessary starting point to embed the objectives of DataPLANT right from the beginning into new projects and proposals. It will help community members on guided data collection and curation as well as on the recording of the complete research context. Building on the successful Galaxy and ELIXIR education services and fully established training courses and channels allows the development of new training programs for the designated community in data and workflow standards/management, data literacy, scientific data analysis, and computational methods. The programs will evolve in the context of the to-be-developed specifications and infrastructures in DataPLANT. This includes both education on the application of the objectives to a bioinformatic research infrastructure and the dissemination of the developed standards, software, and infrastructures into the wider scientific community. The programme will partner with the relevant international communities, like the Carpentries, GOBLET or the ELIXIR Training platform.

Minimum services provided by DataPLANT to foster community acceptance and use. The services defined by community interaction are intrinsically interlinked and depend on each other. The largest drawback of previous efforts did not close the gap between users and services, therefore data stewards in our opinion are the most essential part. The DataPLANT Hub plays an important role to ensure FAIRness for stored data sets and in the consolidation of RDM knowledge within the wider community. The direct provisioning of workflows and storage capacity empowers users to deal directly with their data in a FAIR manner and to incentivize adoption. A couple of measures and the direct interaction with the NFDI and other consortia will provide **contingency measures to ensure permanent availability of services and data security.** The structures we envision are intended to revolutionize research data management for the plant research community for the long-term. Beyond the limited duration of DataPLANT, we see the following sustainability models for the individual components of DataPLANT's strategy, to be investigated in detail during DataPLANT's duration:

- Data stewards and the DataPLANT Hub play an essential role in assisting the user community in regard to adopting and sustaining DataPLANT's RDM standards and practices. We anticipate that in the long-term, large user groups will contribute independent funding towards sustaining such personnel and services. Shaping an understanding – and documentation – of these is a primary goal of DataPLANT, such that

the need for corresponding personnel can be effectively communicated to funding agencies.

- Due to its strong focus on interoperability, DataPLANT's practices and standards will be usable on external resources with little overhead. Thus, if the storage and compute infrastructure services offered within DataPLANT cannot keep up with the growth of data after the initial DataPLANT funding phase, other resources can be used for the same purpose.

- All intellectual property generated towards realizing DataPLANT, especially all standards and ontologies, practices, software implementations, will be published together with comprehensive documentation under an open source license. DataPLANT seek to engage with all interested users and developers within DataPLANT's scope, to supplement and broaden our own efforts; these could be carried on independently, given enough interest from the plant research community.

- We will work towards a strong international plant community, sharing a vision of FAIR research data and software. Community meetings, Collaboration fests, a mentoring program and joint meetings with the ELIXIR plant community will sustain the ideas and services developed in DataPLANT.

The organisational and governance structures of DataPLANT support the continuous development of services and community participation.

4 Work Programme

4.1 Overview of task areas

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 [*Figure 8*]:

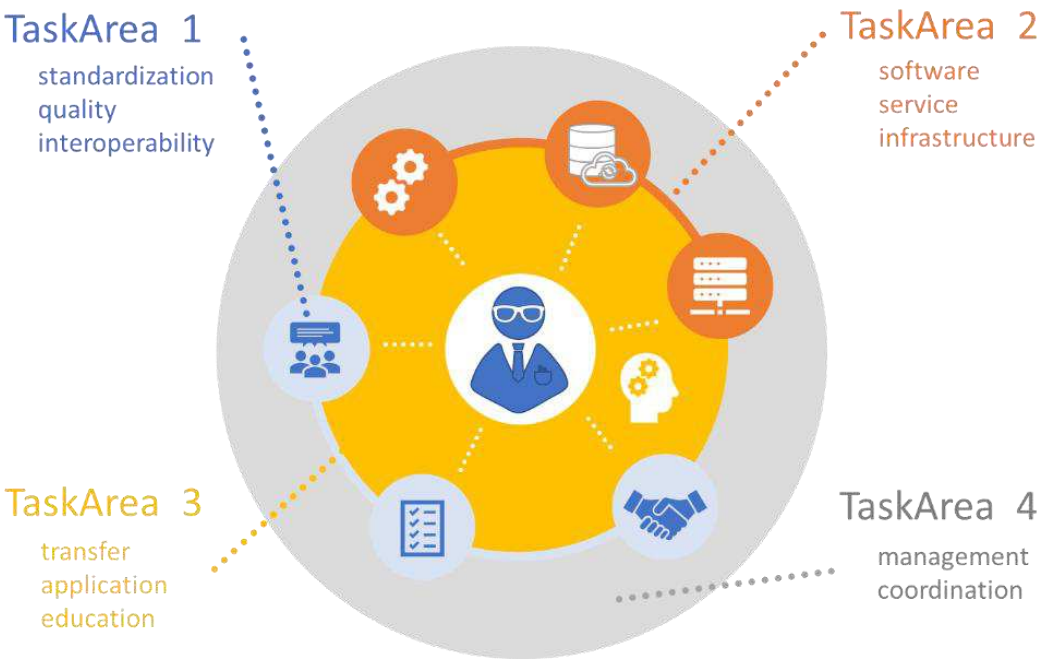

Figure 8 DataPLANT is designed to be user centric. All Task Areas are directed towards the needs of the plant researcher as data champion. The structure ensures the usability in practice and will lead to the formation of a central information resource for fundamental plant research.

Task Area 1 (Standardization, Quality, and Interoperability) will work towards developing the envisioned plant-research (meta)data standards. We believe that for an efficient standardization with respect to ensuring data quality and data/workflow interoperability, an integrative effort is needed that considers these aspects simultaneously through three work packages.

- **Standardization -** DataPLANT builds on a large network of existing co-operations and projects within fundamental plant research to be leveraged to spur plant domain adequate standardization and norms. Thus DataPLANT focuses on the accommodation of the wide variety of necessary metadata and standards within the fundamental plant domain. Uniform standards and procedures as well as a jointly organized and technically distributed data management platform creates added value, both for the DataPLANT community as well as for other disciplines within the whole NFDI. The providers in the consortium can broaden the scope of their services and make use of the knowledge gained from service operation for various communities. Conversely, successful offers for

the DataPLANT community can be transferred to other scientific communities via the NFDI cross-cutting activities. Thus, to foster reusability und long-term access to data sets, the archiving and repository landscape will be evaluated for existing approaches falling back to basic elements from the “Dublin Core” specification as a minimal stop gap solution. DataPLANT will expand and further harmonise existing ontologies and metadata initiatives and/or emerging standards. This includes ontologies, identifiers and interfaces, as well as the establishment of a flexible metadata schema for findability of data sets.

- **Quality -** One of the main aims of DATAPlant is to provide FAIR data and workflows to the community that provides an added benefit. Hence, data quality and especially metadata completeness are of a high importance to safeguard not only FAIR data principle and access but also to be able to empower the community to mine data and to develop added value services.

- **Interoperability –** to ensure maximal data (re)usability and to allow for meta-analysis and data aggregation, interoperability is a major issue that will be tackled by DataPLANT. Thus, DataPLANT will build on existing infrastructure providing unique identifiers, authorization and workflows where possible and re-use extant and accepted data formats. In addition, DATAPlant will together with its user base collaborate with third party providers to improve future data standards, services and workflows.

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, iPLANT 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. The Task Area will provide improvements to data and workflow management across the entire lifecycle of plant research (meta)data. The technical implementation of the NFDI DataPLANT is organized through this task area. The DataPLANT consortium must find answers to the challenges of current and future developments in the field and ensure long-term, productive access to research data. This includes an extension of competencies on all facets of data management as well as the implementation of concepts for sustainable, reproducible scientific methods. The activities in Task Area 2 aim at the provision of software tools, software services, and infrastructure services for (meta)data handling, workflow creation, research data management, and sharing, and knowledge evolution forming the basis for collaborative plant research. These work packages will provide improvements to data and workflow management across the entire lifecycle of plant research

(meta)data. All activities described in the following sections rely on existing infrastructure brought in by the consortium.

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:

- Provide a faceted support infrastructure combining on demand face-to-face consulting with a broad range of assistive services. We follow a holistic approach addressing several target groups including legal advice.
- Building on the successful Galaxy and ELIXIR 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.
- Comprehensive training of the plant research community through workshops and summer schools and providing open training material to support a guided data collection and curation and the recording of the complete data context.
- Application of the objectives to a bioinformatic research infrastructure.
- Dissemination of the developed standards, software, and infrastructures at and beyond participating research centres through partnering in international communities.

DataPLANT is designed to be user-centric, thus it requires specific measures for coordination, consensus seeking and the implemented organisational structures. Data champions and developers should be relieved of administrative tasks to a large extent and be able to concentrate on the implementation of their interests and the coordination of important issues. Thus, central objectives of the task area are:

- Implement and adapt the planned governance and control structures for DataPLANT in order to reconcile the interests and ideas of the community and other stakeholders. There

will be regular evaluations and if required updates of the governance and control structures.

- Together with the general NFDI and the other consortia there will be suitable business and operation models in place for sustainable operation of the identified core services.
- This means developing communication and organisational structures that enable an effective exchange of information between the stakeholders involved and the wider NFDI community.
- At the same time, processes for the comprehensive participation of user groups in corresponding decision-making processes are accompanied during implementation.

The goal is the early and comprehensive integration of all relevant research and interest groups into the processes in order to make strategic decisions and identify possible obstacles or risks at an early stage. In this context, service development with process and business modelling and the integration of external resources is also being promoted. In the overall view of the actors involved, the project positions itself as a specialist centre of bioinformatics around relevant research infrastructures[70]. It moderates the processes necessary for the coordination of all participants by involving the entire NFDI and international structures. This includes the integration of existing or the development of new accounting models, for example in order to map third-party funding flows to resources used.

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

4.2 Task Area 1 (Standardization, Quality, Interoperability)

WP 1.1 Standardization

M1.1.1 Identifiers and Provenance

The use of persistent and unique identifiers such as entity identifiers (e.g. gene, metabolite, protein, reaction etc. identifiers), handles for permanent traceability and identification as well as digital object identifiers (DOIs) is mandatory for archiving and reuse of information objects. DataPLANT will provide data sets in the repository with DOIs using the established Datacite service. At a lower granularity level, research objects will get globally unique identifiers provided by the DataPLANT consortium whereas Open Researcher and Contributor ID (ORCID-iD) will be

used as unique identifiers for researchers. These identifiers are to be used e.g. for the contributor field in the Dublin core specification but also for the reason of crediting data sets unanimously to unique persons. Furthermore, individual entities within datasets will have unique - and potentially versioned identifiers that can be resolved and -if necessary- cross referenced to earlier and later versions and against the EDAM ontlogy[71]. As an example, for a metabolite a unique ChEBI identifier (if it exists)[72] can be used, whereas gene identifiers can disappear or change their meaning with new biological knowledge, thus a unique gene identifier either needs a version prefix (as is common practice for some model plants e.g. Arabidopsis) or it needs to be unique over time as well (e.g. Tomato). Common identifier user practices from the community and relevant other NFDIs will be identified and best practices for DataPLANT will be identified. Together the partners from computer and data science and the plant community based on the feedback and work of the data stewards, will adapt strategies for which raw data, processed data, workflows, etc. which type of identifier are useful and common practices.

Finally, if the data is to be processed through workflows, the underlying workflows will be uniquely identified and linked into the metadata to provide provenance information (detailed in M1.1.3). This measure will be pursued in close exchange with the community and coordinated with other service NFDIs and international communities. In line with this, the focus in this measure is on the clear scientific view, while the technical aspects are addressed in further measures.

Milestones

MS1.1.1.1 Analysis or relevant identifier schemes in other disciplinary NFDIs (Month 3)
MS1.1.1.2 First draft proposal of relevant identifiers and identifier scheme (Month 9)
MS1.1.1.3 Harmonization with other NFDIs regarding identifiers relevant for multiple disciplines (Month 18)

Deliverables

D1.1.1.1 Technical whitepaper describing the first working version of identifier schemes (Month 12)
D1.1.1.2 Updated technical whitepaper describing identifier schemes (Month 24)
D1.1.1.3 Updated technical whitepaper describing identifier schemes (Month 50)
D1.1.1.4 Final technical whitepaper describing identifier schemes (Month 60)

M1.1.2 Metadata standardization and development (Provenance)

The collection of data and their processing needs to completely document the entire data life cycle. Many of the technical metadata required for this life cycle annotation process can be collected (semi)automatically, such as the resolution of the mass spectrometer used, or the precise versioning and parameters of an evaluation tool for ‘omics data and these are often firmly

fixed in globally accepted minimal standards. However, the relevant metadata can change depending on ‘omics discipline and platform used and to fully understand and reproduce a plant experiment more recommended or optional data is often necessary. For this reason, DataPLANT will create a catalogue of metadata that can (or must) be collected for the processing of ‘omics data, relying on i) the user base facilitated by the data stewards and ii) national and international collaborations and initiatives including other potential NFDIs (such as NFDI4Chem, NFDI4AGRI etc.). Required metadata standards will build on the general MIAPPE (Minimal Information About a Plant Phenotyping Experiment) standardization efforts, as the MIAPPE steering committee understands that ‘omics data analysis is a way of molecular phenotyping. That said, due to the rapid progress in bioinformatics, uniform standards, conventions and ontologies must be defined and continuously be updated to reflect technical and scientific developments. Particularly in the comparatively young field of plant ‘omics, such standards and ontologies have not yet been comprehensively established. In DataPLANT, existing standards for ’omics data are to be taken up, expanded and established in the breadth of the community. Together with the technical metadata relevant and useful ontologies such as “Plant Ontology”[73], “Plant Trait Ontology”[74] will be recorded and these will be analysed for their general usability and community acceptability. Based on user feedback, DataPLANT has gained the insight that even generally useful ontologies might lack particular terms and that the formal introduction of these terms into an ontology is generally beyond the effort an experimental lab is willing and/or able to take. DataPLANT will thus support this process by gathering needs of the German fundamental plant science community in terms of ontology developments and will serve as single point of contact towards standardization and ontology bodies profiting from the existing expertise of the DataPLANT members. DataPLANT thus envisions that ontology updates can be streamlined and expedited. The extended ontologies and standards will also help in formalizing metadata about experimental descriptions and pave the way for more machine readable electronic notebooks. Besides specification on metadata and ontologies specific goals of Measure 1.1.2 are the definition and implementation of "data package" standards. Whilst for several data sets, this will rely on final repositories of raw data sets that are mandated both by DataPLANTS and by general community practice (e.g. EBI:ENA for nucleotides, EBI:Metabolights[49] for metabolomics experiments and PRIDE[75] for proteomics experiments) this does not necessarily encompass all (plant specific and relevant) metadata and additional specialized disciplines and/or experiments are usually not represented by specific repositories yet. DataPLANT will heavily rely on the ISA-Tab and ISA-Tools as i) these are proposed by MIAPPE and ii) these best reflect a typical experimentalist workflow and thus would require least work on the users’ side which is an important prerequisite for community uptake and acceptance. On the backend however this is to be represented by Research Data objects (see M2.1.1 for implementation details).

Milestones

MS1.1.2.1 Collection and a formal definition of current relevant standards (Month 4)
MS1.1.2.2 Workshop of the fundamental plant science community requesting comments on MIAPPE (Month 12)
MS1.1.2.3 Identification of bottlenecks and needs of the German community in terms of ontology development (Month 12)
MS1.1.2.4 Identification of lack of annotations in raw data repositories (Month 12)
MS1.1.2.5 First improvements of ontologies (Month 24)

Deliverables

D1.1.2.1 Technical whitepaper describing relevant standards in the fundamental plant community to be published in e.g. F1000 (Month 16)
D1.1.2.2 Definition of a working process to improve third party ontologies (Month 18)
D1.1.2.3 Upstream ontologies usable for a majority for the use cases of the German fundamental plant sciences (Month 50)

M1.1.3 Workflow annotation (Provenance)

In the bioinformatics community and other disciplines where processing is performed with different tools, the term 'pipeline' is commonly used to describe a sequence of individual steps that ultimately lead to a scientific result. The logical consequence is the routine use of reusable, well documented workflows that document the flow of the pipeline and increase the reproducibility of results. Within DataPLANT, standardized workflows and necessary metadata annotations for processing 'omics data are to be established and extended across sites. Whilst DataPLANT will be based on the established Galaxy platform, already support the EDAM ontology[71], DataPLANT together with its user base will identify gaps in annotation in the plant specific domain and to increase versatile reproducibility. The final aim is to map elementary 'omics analysis steps and to offer them across locations. All individual steps should be documented transparently and reproducibly using the annotation with metadata which will be detailed in technical specifications by DataPLANT allowing to also set up new "DataPLANT compliant" Galaxy instances.

Milestones

MS1.1.3.1 Detailed mechanism for workflow annotation specified (Month 16)

Deliverables

D1.1.3.1 Technical specification for workflow annotation and check for reproducibility (Month 32)

WP 1.2 Quality

M1.2.1 Metadata completeness

Metadata completeness will be benchmarked against metadata standards (M1.1.2) and identifiers to be used (M1.1.1). Initially, this will rely on the extant minimal standards (see also M1.1.2), but these will likely evolve to include more recommended and optional data to be included - driven by the needs of the DataPLANT user community. DataPLANT will provide an automatic checking service against three levels: (i) the minimal necessary metadata, (ii) the available recommended metadata and (iii) the optionally available metadata. In addition, DataPLANT, will automatically extract the used and relevant ontologies from individual data sets. As the recommendations and standards are evolving, DataPLANT data sets will be automatically re-assessed on a regular basis and a simple quality indicator such as a traffic light will be provided where red indicates missing metadata, and green would indicate full necessary and recommended metadata. A detailed quality report, will reveal compliance against different standards. The quality re-assessment would also flag Data Stewards and DataPLANT to potentially retrieve additional metadata from already curated data sets, where this is necessary, when new additional fields need to be added.

Milestones

MS1.2.1.1 Definition of a first version of a formal metadata benchmarking set (Month 16)
MS1.2.1.2 Definition of an updated metadata benchmarking set and a clear definition of required, recommended and optional fields (Month 32)

Deliverables

D1.2.1.1 First definition of metadata check procedure (Month 36)
D1.2.1.2 Updated definition of metadata check procedure (Month 40)
D1.2.1.3 Final definition of metadata check procedure (Month 60)

M1.2.2 Raw data (measurement) quality

Besides metadata, DataPLANT will also assess raw data quality. Firstly, users and stewards can provide an optional data field for “perceived, subjective” data quality. This reflects the assessment of the data champions and might e.g. be based on electronic lab books or based on experience of the data handler. DataPLANT will collect data of “lower perceived quality” anyway, as this data could represent meaningful information, especially if inadvertently introduced experimental factors have been recorded in the metadata set (e.g. an unwanted pathogen infection in a drought stress experiment). In addition, DataPLANT will provide quantitative and qualitative summaries of data sets. These would comprise measurements providing information about data sets in a

quantitative way such as size in MB, average number of replicates, and number of biological entities as well as specific discipline summaries e.g. number of reads in an RNASeq experiment, or number of metabolites/proteins measured in metabolomics /proteomics experiments. Furthermore, with a growing data body and based on the experimental condition and sample metadata, DataPLANT will use statistical assessments of data sets against the whole DataPLANT compendium to identify individual data sets that behave differently and/or have individual values that are outside of the typical range. In later stages of the project and by bringing in hand curated legacy data sets, DataPLANT will use machine learning and AI techniques to further predict outlying data sets and variables. This is of particular importance for the plant scientist, as this might flag potential problems in the data sets or it might actually provide insights into biological peculiarities. Thus, this provides a unique advantage of DataPLANT.

Milestones

MS1.2.2.1 Definition of summary data to be produced (Month 8)
MS1.2.2.2 Definition of a best template for subjective data quality (Month 16)
MS1.2.2.2 Prototype of data quality checker based on legacy data (Month 32)

Deliverables

D1.2.2.1 Final definition of necessary data summaries and statistics (Month 16)
D1.2.2.2 Statistical and machine learning data checking procedure (Month 40)
D1.2.2.3 Improved and Updated statistical and machine learning data checking procedure (Month 60)

M1.2.3 Workflow quality and reproducibility (service / executability / workflow compiler)

A major topic in the experimental sciences is data reproducibility, in the ‘omics field this is exacerbated through changing pipelines and non-versioned or non-executing workflows etc. For this reason, DataPLANT will host quality control data sets representing well defined subdiscipline entities together with the expected workflow outputs. These will be sporadically subjected to DataPLANT provided workflows (M2.2.5) and thus check for (i) service executability and by comparing to the expected outputs (ii) reproducibility can be checked. Besides checking for byte identity of resulting datasets, DataPLANT will also check for outcome reproducibility. This is especially important in cases (i) where output is not deterministic and (ii) where subsequent tool versions fix problems in the data workflow. This is particularly important in next generation sequencing analysis where different versions of e.g. the highly used edgeR[76] package produce slightly different numbers of differentially expressed genes, which is both accepted and known in the community but highlights the importance of versioning. Whilst DataPLANT will keep legacy

versions of tools, in cases where third party tools have reported mitigated problems it is necessary to check whether results were due to these problems, or whether evaluated data is still mostly unchanged. To also inform the community of the former DataPLANT will allow automated feedback to users and data champions informing them when new toolchains provide different results or when toolchains might have used non optimal standard parameters. This will build on Galaxy facilities which have proven itself in tracking issues in “standard BLAST“ usage[77]. DataPLANT will however not judge the results.

Milestones

MS1.2.3.1 Prototype of test data sets (Month 8)
MS1.2.3.2 Identification of useful similarity measures for non-deterministic and updated workflows (Month 16)

Deliverables

D1.2.3.1 Definition of testing procedures for deterministic workflow outcomes (Month 16)
D1.2.3.2 Definition of quality measures for non-deterministic workflows (Month 38)

WP 1.3 Interoperability (Provenance)

M1.3.1 data formats

In the area of data management of research data, a particular challenge is to select an appropriate set of (exchange) data formats. Software tools and devices often have their own formats, some documented, some proprietary. Some tools integrate evaluation and visualization environments. This also applies to ‘omics data, where a variety of data formats are used. The special feature of data formats in research is that they are usually task-based and tool-based, but are not necessarily developed against the background of the best possible reusability, archivability and structural transparency beyond the generating software. To the implementation of the FAIR principles, it is therefore necessary to create adapters, converters and parsers, as well as tools for syntactic testing within the framework of quality assurance processes. This also includes a detailed description of the semantics of the data formats and integration into the EDAM ontology. Since many tools already include different (metadata) information for technical reasons, it is necessary to make these metadata available, for example, for detection systems, searches and other software. Since this cannot be done manually for large amounts of research data generated in the same research context, it is necessary to extract the metadata automatically. For this purpose, appropriate tools must be developed for the data used in the project, whereby the extensibility to further tools can be integrated into the architecture of the metadata extractors.

Milestones

MS1.3.1.1 List of raw data formats compiled (Month 8)
MS1.3.1.2 List of semantic qualities and parsers needed (Month 16)

Deliverables

D1.3.1.1 Technical report on relevant data sets in the fundamental plant sciences and additional requirements (Month 20)
D1.3.1.2 Updated technical report on relevant data sets in the fundamental plant sciences and additional requirements (Month 40)

Measure 1.3.2 Interoperability and Cooperation with international Repositories

DataPLANT uses the extant and community accepted minimal measures and data standards as well as formats that are being used or being promoted as standards in general. As such all data sets that have been annotated fulfil the minimal required DataPLANT level (status yellow or above, M1.1.2) including the requirements for third party databases. This is of particular importance as primary data will also be submitted to these databases upon complete annotation and analysis. In particular these comprise the European nucleotide archive ENA[78] (for sequencing data), Metabolights[49] (metabolomics) and PRIDE[75] (proteomics). To keep abreast with new developments and to also improve minimal standards required by these databases DataPLANT will actively collaborate in the extension of these formats. In addition, other experimental plant data not falling into these disciplines can mostly be subsumed as a “plant phenotyping experiment“ in the widest sense. Due to the inherent heterogeneity of these sets, these data are either stored in catch-all databases such as Zenodo or in plant specific repositories such as the e!DAL[79] repositories which provide MIAPPE compliance. Thus, the latter can be fully DataPLANT interoperable based on their metadata and based on the standardized APIs and DOIs featured by these and DataPLANT will actively collaborate with these to ensure metadata compliance.

Milestones

MS1.3.2.1 Analysis of third-party repository requirements (Month 8)
MS1.3.2.2 Liaison with third party archives e.g. EBI (Month 16)
MS1.3.2.2 Meeting/jamboree with third party archives e.g. EBI for future metadata plans (Month 40)

Deliverables

D1.3.2.1 Joined future metadata specification building (Month 50)

M1.3.3 Workflows

DataPLANT will mostly rely on the widespread Galaxy platform[3] and allow Nextflow workflows[80] (M2.1.3). Due to its versatility this allows a “plug” and “play” architecture within DataPLANT to be able to interface with external providers through connectors (M2.1.3). However, DataPLANT will check whether different workflows would be needed in the plant community and recommend Galaxy and/or Nextflow substitutions and/or wrappers where necessary. Based on the current user surveys however GALAXY and Nextflow should be able to tackle (almost) all workflow needs.

Milestones

MS1.3.3.1 Analysis of workflow engines used in the community (Month 24)

Deliverables

D1.3.3.1 Final workflow analysis and GALAXY/Nextflow wrappers suggested (Month 40)

M1.3.4 Infrastructure

DataPLANT will reply on mostly standardized and resilient infrastructures that have been tested and approved in e.g. the established and running Baden-Würtemberg and de.NBI Clouds. As such DataPLANT will build on standard commodity hardware with standardized open source software stacks to facilitate reuse (WP 2.1). In addition, where possible, DataPLANT relies on already established international infrastructure standards and/or brokerage. Examples include the ORCiD service but also the ELIXIR AAI (M2.1.5). This increases usability and reduces efforts on the data plant side. In addition, DataPLANT facilitates and adapts to community accepted raw data storage providers such as the EBI, where this is necessary and provides metadata subsets which are fully compliant with these storage providers. However, DataPLANT also provides additional metadata layers and allows to directly access evaluated and derived data sets.

Milestones

MS1.3.4.1 Definition of interfaces and generalized standards to be used in infrastructure (Month 8)
MS1.3.4.2 User Survey about third party service integration (Month 20)
MS1.3.4.3 Updated definitions of identifiers and services to be used in DataPLANT infrastructure (Month 40)

Deliverables

D1.3.4.1 A comprehensive definition of best practices for third party service integration (Month 50)

M1.3.5 Electronic Lab Notebook

Currently, multiple initiatives and commercial providers are trying to establish electronic Lab notebooks and/or journals, however as of yet no accepted standard platform(s) have emerged. Therefore, DataPLANT will follow open initiatives and participate as a major stakeholder to safeguard adequate use of ontologies and identifiers in electronic lab notebooks, in particular if this is to be done by other NFDIs (e.g. NFDI4CHEM). Also, if open APIs and or standards for electronic lab notebooks arise, DataPLANT will facilitate their inclusion in the DataPLANT. Possible approaches will be discussed with the designated community and interfaces defined for development, if necessary.

Milestones

MS1.3.4.1 User Survey about third Electronic Lab Notebook use and requirements (Month 10)
MS1.3.4.2 User Workshop to define Electronic Lab Notebook requirements (Month 36)
MS1.3.4.3 Analysis of the Electronic Lab Notebook landscape (Month 48)

Deliverables

D1.3.4.1 User Guide to decide on Electronic Lab Notebook use (Month 55)

4.3 Task Area 2 (Software, Service, Infrastructure)

WP 2.1 Software

M2.1.1 Indexing, ontologies, search

The standards and definition arising from Task Area 1 need to be collected, processed and stored in an efficient and reliable way. We envision the usage of the Research Objects Model as general technical vehicle for metadata handling. The ontologies describing the data model will be stored using the RDF format. As the ISA-TAB model can be considered as *de facto* standard in the plant community, it will be used as a template for the input and submission interface. It is desired to store the metadata content in flat files along with the data, efficient converters and adapters are required, to transfer the ISA-TAB based input to e.g. JSON files. The content of the flat metadata files will be made available via a central database, making its content searchable through the DataPLANT Hub (see M2.2.2). The metadata content describing a research project may change and evolve over time e.g. through availability of further time points or additional replicas. To keep track of this evolving information content versioning mechanism and a fully descriptive documentation needs to be available. This is not only true for the evolution of metadata content but also for the evolution of ontologies. Of course, new version of ontologies require a thorough review process and well-defined update procedures. The actual implementation of these pieces of software will be made available through the DataPLANT Hub (see M2.3.3), so that the referenced data and discoveries are reproducible and reusable, fostering IT-based scientific insight.

Milestones

MS2.1.1.1 Implementation of the metadata scheme (Month 12)
MS2.1.1.2 Implementation of converters and adapters (Month 24)
MS2.1.1.3 Comprehensive metadata and ontology versioning (Month 40)
MS2.1.1.4 Identifier for data objects (Month 30)

Deliverables

D2.1.1.1 Full integration of metadata handling through the DataPLANT Hub (Month 50)
D2.1.1.2 Organisational process for reviewing data objects and their annotation (Month 30)

M2.1.2 Cross-linking of information and machine learning

Research data in general and plant research data in particular can only unfold its full information content when different bits of information are contextualized e.g., the up or down regulation of the

expression of a certain protein is a scientific insight just by itself. In order to understand the broader picture, the relevant gene and environmental factors need to be known. This measure aims at the cross-linking of information on the level of the metadata annotated to individual data sets. As it is planned to store metadata relying on RDF (see M2.1.1) the inherent logic is represented as a graph. It is anticipated to use shape expressions for these graphs to formally describe RDF metadata and to automate validation thereof. ShEX[81] will be implemented for defining and validating data records which will be accessible through the DataPLANT Hub.

The wealth of information which will be available through DataPLANT, in particular the logic represented through the cross-linked RDF graphs, will serve as basis for the application of modern machine learning techniques. It is anticipated to identify so far unknown patterns and dependencies. The outcome will serve trivial convenience purposes such as the automatic suggestion of likely attributes when filling an ISA-TAB but also highly complex insight such as hidden regulatory pathways. Consequently, DataPLANT will provide linked open data objects boosting scientific insight for the plant research community.

Milestones

MS2.1.2.1 Implementation of RDF representation (Month 12)
MS2.1.2.2 Cross-linking of RDF graphs (Month 24)
MS2.1.2.3 Machine learning based pattern recognition (Month 50)

Deliverables

D2.1.2.1 Implementation of a knowledge-based suggestion mode (Month 56)

M2.1.3 Workflows and orchestration

The handling of plant research data, its annotation with metadata, quality control and consequently the creation of well-annotated research data objects makes the usage of workflows indispensable. Besides the obvious need for compute (see M2.3.1) and storage (see M2.3.2) capabilities, a well-structured workflow approach is required. DataPLANT will mainly rely on Galaxy, the currently most wide-spread workflow solution in life sciences. The DataPLANT Hub (see M2.3.3) will enable the execution of workflows on resources contributed by consortia members. Further, it will mediate the orchestration of workflows, or parts of them via resources hosted by third parties such as de.NBI Cloud or EOSC-life. An essential aspect is the accurate description and annotation of workflows (see M1.1.3) which will enable their reusability and largely improve the reproducibility. As workflow standards vary over time and among sub-community, DataPLANT will ensure generic and interchangeable connectors enabling the capability for integration and the usage of further workflow languages such as Nextflow[80] promoted through nf-

core. Consequently, users may choose from a set of well-curated workflows most suitable to answer their research question at hand.

Milestones

MS2.1.3.1 Basic implementation of Galaxy workflow execution (Month 18)
MS2.1.3.2 Workflow orchestration over multiple sites (Month 30)
MS2.1.3.3 Basic implementation of Nextflow workflow execution (Month 40)

Deliverables

D2.1.3.1 Full availability of Galaxy and Nextflow workflows (Month 50)

M2.1.4 Portal and portlets

The DataPLANT Hub is at the centre of all infrastructure ambitions. It will serve as the central portal for accessing all kinds of data, services and metadata. Based on an initial evaluation a suitable portal framework (e.g. Flask, Ajax, Hubzero, ...) will be determined. It will form the basis on which different containerized microservices will reside offering different functionalities. These later portlets typically have a connector to the relevant infrastructure service e.g., storage, compute or metadata index and an interface for user input/control.

DataPLANT will collect the requirements of its community for which portlets and underlying services will be developed. It is a clear design decision to follow a decoupled approach in the best sense of modern software engineering. This allows for flexibility and readjustments on short time scales to suit the needs of the community. The collection of portlets will comprise input, upload, search and manipulation portlets. Among the most important ones is the portlet featuring the input mask for metadata input. It requires compatibility to ISA-TAB, connection to the storage infrastructure and efficient access to the metadata registry. Most important are the high requirements with respect to its usability to ensure a high user satisfaction. Further portlets will feature complex search functionality as well as compute capabilities through the execution of workflows on selected data sets. The portlet development is subject to constant evolution. In any case a common look and feel for all user interfaces will be ensured to improve usability and user acceptance. The DataPLANT Hub will also serve as a platform for hosting documentation and training material. There will be an internal section for e.g. technical documentation addressing primarily the data stewards and a public section for the user community.

Milestones

MS2.1.4.1 Definition of technical requirements and selection of software framework (Month 6)
MS2.1.4.2 Base implementation and first portlet (Month 12)
MS2.1.4.3 Organisational scheme for hosting and maintaining documentation (Month 18)

Deliverables

D2.1.4.1 Fully operational DataPLANT Hub (Month 60)

M2.1.5 Authentication and authorization

The authentication and authorization to all DataPLANT related resources will be handled through the ELIXIR AAI (cf M1.1.1 and M1.3.4) building on the Perun identity and access management software. Users have to register for an ELIXIR ID which serves as unique user identification, bridging the gap arising from non-standardized credentials issued by the home organisation of the users and the need for a globally unique identifier for each natural person accessing DataPLANT services. It is anticipated to link the ELIXIR IDs and ORCIDs for identification and credit-related purposes. ELIXIR AAI in conjunction with Perun enables the implementation of a fine granular, role-based model allowing sensible access control to data and services. The consortium can build on practical experience with such implementation arising from de.NBI Cloud and EOSC-life.

Milestones

MS2.1.5.1 Integration of ELIXIR AAI (Month 12)
MS2.1.5.2 Linking of ELIXIR AAI and ORCID (Month 18)

Deliverables

D2.1.5.1 Implementation of fine-grained access model (Month 30)

WP 2.2 Service

M2.2.1 Data staging and handling

The general anticipation for the flow of DataPLANT data assumes the generation of raw data through a sequencer, by a biochemical assay or other means, leading to a more or less structured data set. The obligation of the scientists includes to structure the data and annotate it appropriately while receiving assistance of a data steward. Then the annotated data set is

transferred to the DataPLANT Hub. For this purpose, an upload service will be implemented. Using state-of-the-art transfer protocols, the data is moved to storage resources such as the bwSFS and registered with the DataPLANT Hub. Which transfer protocols are going to be used is subject to evaluation. For sporadic transfers of limited size classic approaches building on rsync or rclone might be completely sufficient, for huge data sets or the integration of remote repositories UDP-based protocols might be used. It is also anticipated to reference third party repositories. The required standardization will be established based on the work of TA1. The staging of data between the DataPLANT Hub, storage resources and compute instances will rely on the same technologies as the upload. The annotation service will provide an interface for the scientists to enter and modify metadata information. It will be closely coupled with the upload service and linked to the metadata registry (M2.2.2). It will also provide connectors to enable bulk upload of e.g. ISA-TAB files and a remote API access. The metadata information will be stored along with the data using the JSON format while its content will be made available via the metadata registry (M2.2.3). The download of annotated data sets will be established through an independent service as it not only requires access to the data itself but also has to be connected to the metadata registry (M2.2.2) and search service (M2.2.4). It will rely on the same technologies as the other data handling services described before.

Milestones

MS2.2.1.1 Implementation of up- and download services (Month 12)
MS2.2.1.2 Implementation of annotation service (Month 24)

Deliverables

D2.2.1.1 Full integration of data handling through the DataPLANT Hub (Month 48)

M2.2.2 Research object as a service

The services of this measure represent the core features of the DataPLANT Hub. When filling the metadata information for a given data set through the annotation service (M2.2.1) a plenitude of values is available. To make this manageable for researchers, a recommender service will suggest likely terms and values. The suggestions will be based on similarity following a nearest neighbour implementation in combination with a plain rule set e.g. specific loci are only suggested if they are present for the initially selected species.

The standards and conventions developed in TA1 do not only need an implementation on a mere software level (WP2.1) but also a certain level of quality control. The metadata standard compliance service will ensure adherence to the ontologies and enforce the presence of minimal set of base metadata information. Furthermore, violations of schemata and the repetitive entry of

non-existing terms will be recorded. This collection of misbehaviour of the service (from a user's perspective) will facilitate the optimization of the service interface and help to improve the metadata schemata. It is envisioned to establish a semi-automatic procedure involving data stewards and scientists to ease the expansion of term schemata as far as possible.

As ISA-TAB represents a *de facto* standard within the community a dedicated base profile generation service will be implemented. It will enable the generation of DataPLANT metadata standard compliant ISA-TAB file generation. The availability of these base profiles shall lower the entry barrier for novice users. On the one hand they still may use the things they used to while on the other hand the resulting ISA-TAB files will be fully DataPLANT compliant.

Milestones

MS2.2.2.1 Recommender service for ontology-based annotation terms (Month 24)
MS2.2.2.2 Metadata standard compliance (Month 24)
MS2.2.2.3 Base profile generation (ISA-TAB templates) (Month 36)
MS2.2.2.4 Term schemata expansion (Month 36)

Deliverables

D2.2.2.1 Full research object handling capabilities through the DataPLANT Hub (Month 60)

M2.2.3 Crawling, indexing, metadata registry

The metadata information describing a research object is stored along the object using plain JSON files (M2.2.1). At the same time the central metadata registry plays an essential role to make research objects findable (M2.2.4). Hence the federated approach for storing the plain metadata in a flexible way needs to be connected with an approach best described as monolithic data warehouse. For this purpose, it is planned to develop a metadata crawler service which works its way over the indexed metadata, capable to find and interpret so far unregistered metadata. In such a way the hard to fulfil requirements for metadata completeness are handled in a way that allows for data and metadata growth and extension without breaking schemata and conventions.

In a similar way metadata information describing workflows will be handled. In contrast to metadata JSON files, it is planned to store workflows in a GitHub/Gitlab-like environment allowing indexing and versioning (M2.2.7). This facilitates the crawling for new information which will lead to a workflow index enriched with EDAM ontologies.

Milestones

MS2.2.3.1 Implementation of metadata crawler (Month 12)
MS2.2.3.2 Indexing of metadata content (Month 18)
MS2.2.3.3 Implementation of central metadata registry (Month 24)

Deliverables

D2.2.3.1 Fully operational metadata index (Month 30)

M2.2.4 Search

Findability of data and supplementary materials by persistent identifier is a key requirement named first in the FAIR principles. In order to perform data integration and reveal additional knowledge, searching and parsing of individual experiments according to keywords and metadata is a valid approach. Therefore, we will implement Elasticsearch on our research object data stored in JSON. Elasticsearch provides a distributed, multitenant-capable search with a web interface to enable findability of all data. However, the task of integrating the cross-omics data sets is inherently complex and challenging, because the individual entities within the data set are highly interconnected. To access this interconnectedness and having the possibility to retrieving subsets by specific properties and metadata annotations would dramatically facilitate the data retrieval for integrated knowledge discovery. The data layer follows a relational data model with metadata stored in direct JSON format. Apart from using the in-built indexing features of the database system, we use existence indexes (Bloom filters) for faster query processing and to incrementally explore the data. Equipped with a new query language to incrementally explore the data through interactive predicate construction, we provide a way to query biological data more flexible in much deeper levels without any knowledge about the underlying schema that is mostly complicated and highly domain specific. Finally, this measure will ensure the search capabilities of the DataPLANT Hub to allow novel insights leveraging existing data.

Milestones

MS2.2.4.1 Implementation of Elasticsearch (Month 28)
MS2.2.4.2 Development of an integrational relational data model based on RO metadata (Month 44)
MS2.2.4.3 Design specialized indexing mechanism and query language (Month 52)
MS2.2.4.3 Implementation of a GUI to access the cross-omics data (integration into DataPLANT HUB) (Month 60)

Deliverables

D2.2.5.1 First operational search capabilities (Month 38)
D2.2.5.1 Fully search capabilities with interactive predicate construction (Month 60)

M2.2.5 Workflow execution, orchestration and long-term executability

The Galaxy workflow framework provides an ideal platform for DataPLANT users and aligns well with the goals of the NFDI. DataPLANT will work closely with the European (and world-wide) Galaxy community, support the development of novel Galaxy workflows for users and make sure that these can be reproducibly executed on different infrastructures (e.g. EOSC, de.NBI Cloud, bwCloud, BinAC, Nemo). To circumvent any scalability issues and make effective use of all different resources in DataPLANT, Galaxy will orchestrate workflows across different infrastructures. Partners with high-memory nodes for example will get jobs with a high demand of memory, whereas other jobs are scheduled to other locations. The job profile that decides where a job gets scheduled will be learned based on previous job execution, so that over time the distribution of jobs is getting more efficient. Galaxy will also serve as a general-purpose workflow execution service that supports the GA4GH WES and TES API. With this is will be able to schedule workflows in a standard way also from other workflow engines.

The European Galaxy server is following an Open Infrastructure model, which means that the entire system is described in machine readable formats, like *terraform scripts*, *ansible playbooks* or *packer definitions*. This will facilitate the integration of Galaxy into the DataPLANT Hub and the provision of the workflow execution and orchestration services.

However, for DataPLANT we intend to go a step further, by freezing every workflow, with its tools, reference data sets into its own container, creating a referenceable bundle. The resulting service guarantees the preservation of the execution environment.

Milestones

MS2.2.5.1 Workflow execution service (Month 30)
MS2.2.5.2 Workflow orchestration service (Month 42)
MS2.2.5.3 Reliable freezing of execution environments (Month 48)

Deliverables

D2.2.5.1 Fully operational workflow services (Month 50)

M2.2.6 Interactive data visualization

In DataPLANT we will focus on the visualization of tabular and graph-based data which constitutes a majority of data available in the biology context. To integrate guidance in DataPLANT, we first identify best practices and frequently used techniques. A baseline are existing surveys on biovisualization techniques[82–84] and augment those by additional techniques using surveying techniques proposed for the biology context[85]. In most cases the same data can be visualized using various techniques and workflows or variants of workflows. This information is best communicated visually with a workflow browser that graphically depicts steps of workflows, alternatives, requirements (e.g. data quality, input). Here we implement such a browser that integrates with ontologies to automatically suggest possible workflows for a selected set of data items.

Many visualization techniques come with a large set of parameters and interaction possibilities. In order to be able to recreate a visualization, these settings have to be recorded, i.e. we have to keep track of provenance. A number of systems have been proposed in the field of visualization[86–89].This work needs to be reviewed and adapted to the DataPLANT setting. The best practice workflows along with their parameter settings are a template to start from, for the user of DataPLANT. Starting from this template, the best workflow has to be selected and adjusted to the given data e.g. handle missing data elements, cope with varying data quality, cope with varying data sizes and complexities. We will develop a classifier that suggest best settings based on data learned from previous examples. Input features for the classification process can be extracted from the ontologies.

Milestones

MS2.2.6.1 Identification of best practices for context-based visualization (Month 18)
MS2.2.6.2 Development of a visualization and workflow browser (Month 38)
MS2.2.6.3 Integration of visualization provenance (Month 52)
MS2.2.6.4 Implementation of guidance procedures (Month 56)

Deliverables

D2.2.6.1 Full implementation of a visualization and workflow browser (Month 50)
D2.2.6.2 Implementation of a visualization browser including guidance procedures and provenance (Month 60)

M2.2.7 Versioning services

To model and capture the evolution of data (references), metadata, and workflows, we will focus on version control semantics. The evolution of a data context is modelled as a sequence of discrete snapshots in time that represent the history of a context. From this, it is possible to identify differences between versions (cf. M1.1.3), reconstruct context history without gaps, and understand the effect of changes to the context. Moreover, distributed version control allows decentralized storage and evolution of data context and therefore an ideal semantic framework for high-frequency collaboration. For example, simultaneous modification of a context by several collaborating researchers (and, in the context of DataPLANT, data stewards, see M3.2.1) without manual synchronization is an anticipated use case. Furthermore, it is not necessary to operate on a single copy of a data context; rather, multiple copies can be created from any version (branching), evolved independently, and then later merged back into a single context. Distributed version control is recognized as a best practice in collaborative software development, and DataPLANT's goal is to make all the advantages of corresponding technologies available to plant researchers, thereby in particular providing ground-breaking advantages for collaboration and reproducibility.

We will therefore implement a general version control service for data contexts and integrate it with the other services provided in TA2; in particular, integration with measures M2.2.1–M2.2.7 which all interoperate on data contexts. Towards an implementation, we will layer data context versioning atop an existing open-source distributed version control system (likely Git; other options include Darcs or Mercurial) and expose it to researchers using the DataPLANT Hub as platform. Versions will be automatically equipped with unique identifies that can be used for publication of entire data contexts and credit claiming.

Beyond simply recording changes to metadata, changesets between versions allow an automated annotation of metadata history, which is important for searching and indexing (M2.2.3 and M2.2.4) metadata changes. Beyond this, consistency and adherence to DataPLANT's schemas can be checked and enforced at the changeset level (M2.2.2). For workflows, two strategies will be implemented: either include the full workflow definition in the data context, or alternatively, refer to an existing (possibly standardized) workflow in a workflow registry.

Specific attention will be devoted towards automating the merging of diverged data contexts that occur frequently in collaboration; here, dedicated conflict resolution strategies will be developed tailored to data, metadata, and workflows. Further adapting established paradigms to DataPLANT's mission, we will investigate the automated execution of workflows after changes to

data contexts (such that these can include processed data), as well as efficient and comprehensible interfaces for visualizing and browsing data context histories.

Milestones

MS2.2.7.1 Basic versioning services for data contexts (Month 18)
MS2.2.7.2 Dedicated merging strategies for data contexts (Month 26)
MS2.2.7.3 Browsing and visualization of data context histories (Month 36)
MS2.2.7.4 Versioning and automated execution of workflows (Month 42)

Deliverables

D2.2.6.1 Fully operational and integrated services for versioning data, metadata and workflows (Month 46)

M2.2.8 Risk assessment, risk management and preparations for long-term access

Re-using data-sets requires usually a complex software set, e.g. operating system and its configuration, libraries and domain specific software. In some cases, data-sets can be migrated to be used with current software, however, specialisation of research together with a fast technical life-cycle leads to a fast-growing diversity of formats used in research data sets. Over time the availability of this software stack and the ability to run such a software stack will deteriorate and eventually software will be inaccessible and unusable. Furthermore, today's computer assisted research does not only rely on existing software and digital resources, but increasingly devotes significant resources into creating new digital resources and tailored software-based methods, i.e. to process data or to create novel (software-based) models, published independently or together with datasets. The purpose of this measure is to implement infrastructure to ensure access to software, necessary to render, inspect and reuse. As part of the archival (object ingest) workflow, the research object will be analyzed and software dependencies will be identified. Based on the result, risks for future access will be assed. This result will be used for feedback to the creators as well as to keep records of objects to be maintained and curated., To cope with software dependencies and to implement a runtime platform we plan to integrate with the infrastructure from internationally operating software preservation initiatives (EaaSI, Software Heritage[90], UNESCO Persist) and implement a service for publishing long-term accessible scientific software environments.

Milestones

MS2.2.8.1 Risk assessment for standard use cases (Month 18)
MS2.2.8.2 Dependency analysis of execution environments (Month 42)
MS2.2.8.3 Software access workflow service (Month 36)

Deliverables

D2.2.8.1 Feedback service on ingest of future access risks (Month 24)
D2.2.8.2 First prototype of runtime environments access service (Month 30)
D2.2.8.3 Fully operational runtime environments access service (Month 48)

WP 2.3 Infrastructure

M2.3.1 Compute

All services described earlier require an appropriate compute infrastructure which needs to be operated. The requirements can be roughly grouped into three categories: 1. Web services and microservices such as the DataPLANT Hub, annotation and search services, metadata registry and visualization services. 2. Crawling and indexing services, internal machine learning workflows 3. Quality control and research workflows. The first category will reside within virtual machines hosted via the de.NBI Cloud. The computational needs are rather small, but high availability and failover mechanisms are desired to ensure resilient operations of the DataPLANT Hub. The second category refers to internal, not directly user triggered, computations. The computational need is moderate but the vicinity to the DataPLANT Hub and the majority of the stored data would be beneficial. Hence this part will also be mainly handled through de.NBI Cloud resources, complemented with by HPC clusters when needed. The third and last category describes pure computational workload, which in principle could be processed everywhere. We will employ cloud bursting techniques and rely on established techniques already in use by the Galaxy community, or within the ELIXIR context.

All infrastructure and all services residing on it needs to be monitored. We envision a Grafana, InfluxDB, Telegraf stack for monitoring and consequently also for accounting of resource usage.

Milestones

MS2.3.1.1 Connection to cloud-based compute resources (Month 12)
MS2.3.1.2 Remote workflow execution (Month 24)
MS2.3.1.3 Monitoring and accounting (Month 36)

Deliverables

D2.3.1.1 Fully operational compute platform (Month 40)

M2.3.2 Storage

Similar to the compute infrastructure different storage resources are required to make the different DataPLANT services available. Again, a differentiation into three categories seems appropriate: 1. Hot data directly accessible and processable such as upload data, but also intermediate data from internal processes such as crawler temporary files. This kind of data is usually processed instantaneously and typically requires fast storage. Ephemeral storage or alike available through the virtual machines of the de.NBI Cloud shall fulfil this need. 2. Luke warm data may comprise partially processed data sets or data objects with incomplete metadata annotation. Such data shall reside on cinder volumes, cluster file systems or similar resources which can act as intermediate storage for a given time. On the long run it is anticipated to integrate the cache storage of the Storage for Science in a seamless way, allowing a buffered data staging among all sites. 3. The DataPLANT data repositories will serve as long-term archives for published or publishable DataPLANT data in accordance with the FAIR principles. The DataPLANT data stewards will help and ensure the scientific annotation and curation of the data. The repositories content will be stored on the Storage for Science.

It is essential to develop suitable staging mechanisms between different storage resources. This affects all DataPLANT services and the connectivity of the DataPLANT Hub. Monitoring and accounting capabilities will be implemented along with MS2.3.1.3.

Milestones

MS2.3.2.1 Connection to federated storage resources (Month 12)
MS2.3.2.2 Efficient data staging (Month 24)
MS2.3.2.3 Monitoring and accounting (Month 36)

Deliverables

D2.3.2.1 Fully operational storage platform (Month 40)

M2.3.3 Training Infrastructure and WaaS

In an ideal setting trainees can use the same infrastructure as utilized anytime later. This means the training infrastructure should be the production DataPLANT infrastructure. As it is planned to offer the DataPLANT infrastructure as a Workshop-as-a-Service (WaaS), the assignment of sufficient resources to individual workshops needs to be orchestrated. For this purpose, a special training queue will be established, blocking sufficient resources to carry out workshops even with larger groups.

The combination of the developed training material with the DataPLANT WaaS will enable trainers to dramatically cut down the time needed for workshop preparation as a trainers does not need to worry about any maintenance or administration and the training material as well as the used workflows are constantly tested.

Milestones

MS2.3.3.1 Reservation of dedicated training resources (Month 12)
MS2.3.3.2 Testing and hosting of training material (Month 24)

Deliverables

D2.3.3.1 Fully operational training platform (Month 40)

M2.3.4 DataPLANT Hub

The DataPLANT Hub will be the central nexus for all users accessing DataPLANT services. It requires a solid software framework (M2.1.4), connectivity to all services, databases and repositories as well as a reliable compute (M2.3.1) and storage (M2.3.2) infrastructure as basis. Its fundamental operation is at the core of this measure which includes failover mechanisms ensuring a high availability of all services. The interface for the access to all content of DataPLANT will be handled through individual portlets, one for each service. Their look-and-feel will be harmonized in such a way that even inexperienced users may adopt quickly to the use of the DataPLANT Hub. The hub will also orchestrate the workload on at least two different levels. Internal processes such as metadata crawling may generate significant load which shall be offloaded automatically to child instances. External workflow orchestration (MS2.2.5.2) also requires a central component keeping track of all workflows and job traces. All services of DataPLANT require a fine-grained role management and access control. While some information may be accessed by everybody, other bits of information have to be classified till e.g. a patent was granted. The hub will have a uniform rights management covering all portlets ensuring full integrity and data safety. The operation of storage and compute resources will be monitored (MS2.3.1.3 and MS2.3.2.3) including workflow execution on remote resources. The DataPLANT Hub will provide a monitoring and accounting portlet with adjustable view for users and administrators.

Milestones

MS2.3.4.1 Interface design and look-and-feel (Month 12)
MS2.3.4.2 Role management and access control (Month 24)
MS2.3.4.3 Monitoring and accounting interface (Month 36)

Deliverables

D2.3.4.1 Fully operational DataPLANT Hub (Month 60)

4.4 Task Area 3 (Transfer, Application, and Education)

WP 3.1 Transfer

M3.1.1 Interconnect users with DataPLANT Hub infrastructure by data stewards

Data stewards operate at the core of DataPLANT interacting with the community directly in their daily tasks on a regular basis. The transfer of knowledge and information is multi-faceted. While formal training and online materials primarily carry information into the field, the data stewards act in both directions. They collect feedback from the field, like on novel workflows seen in operation or learn about needs unanswered by the actual infrastructure and services or inefficient or failing procedures. Data stewards bridge between single research groups and the developers of the core DataPLANT services and providers of the infrastructures. They facilitate the creation of working groups on standardization. To ensure well annotated research objects, data stewards will actively

participate in metadata annotation of raw data to prepare data sets for sharing and publication. Data stewards foster compliance with respect to standards and conventions developed and provided by DataPLANT. They might act as quality assurance for FAIRness at this point as well. In the final stages of a project, data stewards can be requested to review research data objects for completeness and correct annotation and properly hand over all remaining data and ensure the fulfillment of the compliance regarding the good scientific practice. It is important to mention that the ultimate goal is to enable the researcher to produce well annotated research objects finally without or only minimal support by the data steward and only rely on the technical auxillaration by the DataPLANT Hub. Therefore, we aim for mechanism that awards well annotated research objects (see. M3.2.1 Data steward coordination).

Milestones

MS3.1.1.1 Monitoring the needs and demands of the user community (ongoing)
MS3.1.1.2 First suggestion on a survey to oversee interaction (Month 12)
MS3.1.1.3 Exchange with at least one third of the participants (Month 12)
MS3.1.1.4 Exchange with more than half of the participating users (Month 24)

Deliverables

D3.1.1.1 Surveys on the state interaction (Month 24, 48)
D3.1.1.2 Current data set are integrated into DataPLANT as well annotated data objects
D3.1.1.3 User can use DataPLANT Hub infrastructure and produce well annotated data objects

M3.1.2 User-driven customized workflow integration

The development of specifications and the refinement of standards require timely moderated interaction between the practitioners in the field, the developers and the working group assigned to a specific task. Data Stewards in their hinge position collect the necessary information and aggregate it for take up in TA1. This process repeats in cycles. Data stewards monitoring the needs and demands of a specific user within the community. They will introduce new tools and analysis capabilities to the researcher. Further they will custom fit workflows according to lab specific environments in strong coordination with DataPLANT. Software packages can be adapted and included into Galaxy workflows in the DataPLANT Hub.

Milestones

MS3.1.2.1 Monitoring the needs and demands of the user community (ongoing)
MS3.1.2.1 Exchange workflow relate ‘best practice’ with at least one third of the participants (Month 12)

MS3.1.2.2 Exchange with more than half of the participating users (Month 24)

Deliverables

D3.1.2.1 Report on the of standardization with other consortia (Month 24, 48)
D3.1.2.2 Analysis and processing workflow are anchored in the plant community

M3.1.3 Topic specific information channels and active participation

We will foster user participation on several face-to-face and virtual levels to provide a permanent flow of information between the stakeholders and maintain active participation of all involved participants and users. Step-by-step the DataPLANT Hub will fill its role as main point of contact for all relevant data management and workflow related activities. Depending on the size of the groups within the project (general assembly, committee, working groups, data stewards, developers) various communication and coordination tools will be offered, like slack, tickets, wiki, mailing lists). For the wider scientific community and certain forms of educational materials videos channels will be used as well. Infrastructure to support these channels is provided by TA4.

Milestones

MS3.1.3.1 List of possible communication tools distributed for decision for each type of body (general assembly, committee, working groups, data stewards, developers) (Month 4)
MS3.1.3.2 First description of service catalog for information channels (Month 12)
MS3.1.3.3 Aggregate user groups by subject and demands (Month 16)
MS3.1.3.4 Establishing direct communication and discussion (Month 20)
MS3.1.3.5 Maintain communication, consultation and discussion (ongoing)

Deliverables

D3.1.3.1 Communication tools decided and made available for each type of body (Month 6)
D3.1.3.2 First communication concept and tools of DataPLANT Hub decided (Month 12)
D3.1.3.3 Communication channel for bioinformatic consultation (Month 15)
D3.1.3.4 Communication channel for experimental/methodological consultation (Month 15)
D3.1.3.5 Survey (Month 24)
D3.1.3.6 Reviewed and adapted concept for DataPLANT Hub (Month 36)

M3.1.4 Community wide outreach and dissemination

The responsible and informed handling of research data is part of good scientific practice and is therefore just as much a part of everyday research as the quoting of scientific articles. Therefore, it is a central goal of DataPLANT, in addition to the provision of appropriate infrastructures and workflows, to provide experienced researchers as well as junior scientists with up-to-date information on RDM as early as possible. They should familiarize themselves with workflows and the handling of research data through consulting and training offerings. In the long run such qualification measures should be included in the relevant curricula. The task of the work package is also to prepare tailored content for the various tasks and workflows relating to data management over the entire lifecycle. The workflows, methods and application of research data management should be taught to prospective scientists early in their studies. Here, DataPLANT consciously uses already established offers of its partners such as training courses developed within de.NBI and/or courses developed within the universities. The expertise among the participants will be increased and dissipated. The outreach and dissemination activities will be supported by TA4 general outreach and wider information activities (complemented by M4.2.6).

Milestones

MS3.1.4.1 Structure of information materials

Deliverables

D3.1.4.1 Tailored material on RDM for domain specific courses (Month 24)
D3.1.4.2 Expert network exchange platform

WP 3.2 Application and Consulting

M3.2.1 Data steward coordination

Data stewards are a significant resource which needs to be properly managed and fairly distributed. The group of data stewards maintains a permanent link into the community as they accompany scientists and research groups in the various stages of research. They can be requested in various stages of the project and data life cycle (see M3.2.2). As they can spend significant time with single groups, both timely and efficient scheduling is required. To follow the transparent communication and broad user involvement objectives a balanced fair-share algorithm got created:

1. First time is (automatically) granted but goes with conditions (commitment on NFDI objectives, provisioning of the data to the NFDI)

2. FairShare: Available data stewards hours are divided by participants (plus 30% future participants) (to be refined, to include group size etc.) or “own money” (material costs, part of their accepted grant) and bonus points

3. During phases of higher loads order multiple incoming requests by waiting time. Groups which interacted more lately with a data steward will wait comparably longer than researchers who used their services a longer time ago

4. Evaluation of data quality by the board (check on fulfilled conditions, extra points/hours by providing annotated data of a plant or similar), point-system to create an evaluation metric (if and how much)

5. Award extra points for exemplary data sets published and referenced.

The algorithm combines factors of fair distribution of resources with incentive schemes to improve metadata quality and FAIRness of data sets. Additionally, data stewards need to be regularly qualified (see M3.2.5) to be up to date on recent developments in DataPLANT, evolvements in the field as well as international activities and achievements. Regular meetings need to be coordinated online and in-person supported by the office (see M4.1.1) to form the boards of stewards deciding on users requests and requirements.

Milestones

MS3.2.1.1 Setup of the data steward request granting and coordination board (Month 3)
MS3.2.1.2 Review of the assignment and operation procedure (Month 9)

Deliverables

D3.2.1.1 First evaluation and suggestions for modification of the data steward concept (Month 12)

M3.2.2 Data stewards research process design and planning

Data stewards can be applied for in every stage of a research endeavor involving data management. In the preparation stage data stewards advice on data management and standards (detailed in TA1) related questions of a grant application. They will give an overview on the relevant ontologies and metadata to use, help with the design of workflows to deploy and software to involve and calculate compute and storage resources which are required. From this they can derive which funds need to be planned for the proposed project and how much data will be stored and getting published for the long-term. Further they will suggest data formats to be used. After

project kickoff data stewards will support the group members to properly organise their workflows, access the necessary resources and implement their data management e.g. through on-site workshops. They will answer to specific needs and give feedback to developers and service providers. If necessary, they will help to access remote resources or adapt local software packages to be included into Galaxy workflows. In later phases of the project these activities center around the DataPLANT hub (laid out in M2.3.3).

Milestones

MS3.2.2.1 Resource planning for long-term storage, repositories (Month 12)

Deliverables

D3.2.2.1 Survey on ontologies used (Month 8)
D3.2.2.2 Survey on published data (Month 24)

M3.2.3 Legal advice and support

DataPLANT will provide legal advice either directly or via hired expertise. The legal support will work on the field of authorship of data and workflows and addresses potential data and software copyright issues. As DataPLANT is committed to the idea of Open Science and open data it will anchor these principles for the handling of data in the community, in particular through consulting by the data stewards. Copyright issues might concern both the reuse of data by third parties and the safeguarding of their rights as well as the safeguarding of rights to the results of their own research. The rights must be comprehensively clarified with the main stakeholders from funding agencies, universities and researchers and suitably documented in the metadata. The basis is to be established with the designated scientific community and via cross-cutting topic with other consortia and the general NFDI. An essential factor in hosting research data is the rights associated with it, which must be taken into account by a repository operator. The objective of DataPLANT and the participating researchers is to provide the most open and free access to research data possible and therefore advocate the most open licenses for data and their metadata, such as CC-BY. Community open licenses allow a simple technical implementation, but for each research data record, it must be ensured that the data provider places the data under one of these licenses. Software licenses represent a further challenge: Here it has to be clarified how to deal with commercial packages, especially in the long-term. Access rights even after ending a software license agreements might need to get ensure, or escrow services for such software packages need to be created to allow the rerun of a particular workflow to access data sets or verify results. A software package that is no longer available may be necessary for long-term access in connection with reused data. Appropriate agreements must be made with the

software publishers or developers. For Open Source packages with multiple dependencies, it must be clarified how these affect licensing. Data stewards will act as facilitators for legal advice (see M3.2.2) which might be required during the various stages of the planned project e.g. to clarify licenses of software, the ownership of data set, the handling of sensitive information (e.g. location data of rare species; patentable knowledge;) In addition, especially in the field of plant biotechnology and the 'omics field the Nagoya protocol on Access and Benefit sharing advice and best practices are often necessary. They channel requests from the research groups to the legal advisor. The legal advisor will to group and generalize requests and either answer the problems directly or find answers through NFDI cross-cutting activities, consult general NFDI experts or hire an external expert for that particular problem. Suitable answers and proven solutions will be channelled back into the community and the wider NFDI level.

Milestones

MS3.2.3.1 Aggregation of topics where legal advice is sought for (Month 12)
MS3.2.3.2 Exchange topics of legal advice and support with other NFDIs (Month 24)

Deliverables

D3.2.3.1 Official contact person for legal advice and support (Month 8)
D3.2.3.2 Knowledge base on legal aspects for relevant use cases (Month 36)

M3.2.4 Data stewards capacity building and permanent qualification

To achieve a significantly pervasion of the community the consulting and qualification capacities needs to be extended over time. In a staged process the relevant multipliers will get addressed and qualified by both data stewards and DataPLANT lecturers to take an active role in their groups to spread the knowledge on data management, standards and services. The freshly qualified data management specialists of the individual research groups will receive further regular trainings and qualifications on ongoing developments. Train the trainers - the data stewards take actively part in trainings and workshops held by the DataPLANT lecturers to keep track on all relevant developments. They will attend regular meetings to exchange on best-practices, qualify on new standards, learn on solved legal issues, updates on extended, modified ontologies and metadata schemas as well as on potential new workflow and software options.

Milestones

MS3.2.4.1 Regular training units for data stewards (Month 1,13,25,38,50)
MS32.4.2 Identification and qualification of multipliers at data champions' sites (Month 24, 48)

Deliverables

D3.2.4.1 Highly trained (extended) pool of data stewards that are technical up to date.

WP 3.3 Education

M3.3.1 Inclusion in higher education curricula

This measure focuses on expanding the content of on-site and e-learning courses and broadening the range of topics to be taught for students in graduation. In coordination with the participants and wider plant community, the needs of users of the various plant/bioinformatic research infrastructures and experimentalists will be collected and gradually converted into corresponding course offerings. The measure provides lecturers support in integrating DataPLANT's working methods and systematics into their curriculum. Lecturers are continuously informed about the results of method development and data management and course material is provided. They receive information on suitable data sets, the use of the respective DataPLANT repositories and readymade slides and teaching sections. In this way they can teach their participants the necessary tools as well as suitable data sets for experiments and self-study.

Milestones

MS3.3.1.1 First version of a university module template for integration into teaching plans (Month 6)
MS3.3.1.2 Coordination with the wider NFDI (Month 24)

Deliverables

D3.3.1.1 Teaching templates and materials including module descriptions to be used in courses (Month 12)
D3.3.1.2 References to RDM in curricular of early adopters (Month 26)
D3.3.1.3 References to RDM in curricular of the wider community (Month 46)

M3.3.2 Workshops and training courses

In order to avoid time-consuming post-processing of data records in the context of a publication, it is important to advise researchers at an early stage in the life cycle of data management. The aim is to work towards a well thought-out and structured data preparation from the outset, if possible. Data sets are to be prepared for publication as early as possible, enriched and converted into sustainable file formats. Within the framework of regular small group training courses, at working group level or, in the case of methodologically similar procedures in the specialist community, also across working groups, researchers are fundamentally introduced to

methodological, organisational, technical and legal questions of research data management on the one hand, and specific requests of the working group are dealt with on the other. The group of junior scientists needs an increasing amount of qualified Knowledge to access the various advanced research infrastructures and to properly handle the associated data management. Since a junior researcher is usually actively involved in research projects, corresponding courses should take place at the beginning of the research project if possible, so that the life cycle of the research data can be covered almost completely. As a contact point for the implementation of these events, cooperation with the local continuous qualification institutions of the participant institutions is an obvious option. Further cooperation possibilities exist with the training and summer school activities of Galaxy on a national level and ELIXIR/EOSC on an international level. For researchers who are integrated into the chairs through their research, further individual formats should be developed within the framework of this measure in addition to the implementation of advanced courses, e.g. within the framework of the regular colloquia. Training courses for individual working groups with direct reference to the research data generated there are conceivable here. The qualification of the researchers is to be regarded as particularly important, since these generate on the one hand a lot of data and are responsible for the reusability of these. On the other hand, the researchers in their role as supervisors for students and doctoral students have an exemplary character and should therefore adopt a sustainable approach to Research data, for which this project will provide the necessary infrastructure. At the same time, step-by-step instructions for typical processes and instructions for subsequent use by new research groups will be created on the basis of the developments in TA2.

Milestones

MS3.3.2.1 Development for a first training course material for metadata standards and importance (Month 6)
MS3.3.2.2 Development of omics discipline specific training course materials (Month 16)
MS3.3.2.3 Development of data integration course materials (Month 22)
MS3.3.2.3 Improved training course material based on user feedback and new developments (Month 30)

Deliverables

D3.3.2.1 First training course held (Month 12)
D3.3.2.2 At least 12 training courses conducted (Month 60)

M3.3.3 Data management planning

Project managers and principal investigators setting up a research project or applying for a grant should be qualified in holistic planning. They should be informed on the ongoing activities in

standards development, relevant workflows for their project as well as proper licensing of data and software if applicable (see M3.2.3). Further, storage capacity and compute resources should be estimated and applied for. Depending on the future NFDI financing model appropriate funds should be planned for expected support services. The amount of data sets handed over to long-term storage and access for publication should be estimated and taken into consideration for the financial planning as well. Thus, an introduction to data management plans are a prerequisite of successful grant applications and a project start. The training courses in data management plans represent a preliminary stage to the consultations by the data stewards (laid out in M3.2.2) in the individual case with the aim of clarifying general questions in order to be able to deal in detail with special questions and cases for the individual consultation.

Milestones

MS3.3.2.1 Check best suitable available data management plan tools to be used for DataPLANT (Month 4)

Deliverable

D3.3.3.1 Generated Tutorial how to generate data management plans using extant tools (Month 14)
D3.3.3.1Provide filled in samples, templates for typical use cases, text blocks for grant applications describing the relevant points (Month 36)

M3.3.4 Online training material

For a geographically distributed community online educational materials complements the data stewards in one-to-one consulting, face-to-face workshops and training sessions. These resources will deliver up-to-date information on ongoing standardization (orchestrated in TA1), running and upcoming services (see TA2). The resources will be matched to the different target groups of students, PhDs and operating personnel in the research group. We will embed DataPLANT online materials into local activities of the institutions to align them with the institutional strategy (see M3.3.5) as well. Further we will work towards a common training platform offered in the context of the general NFDI, coordinated with the other consortia. DataPLANT will use the Galaxy training-as-a-service platform to provide easy-to-use on demand resources within its own infrastructures.

Milestones

MS3.3.2.1 Development for a first training online course material for metadata standards and importance based on in person training course (Month 14)

MS3.3.2.2 Development of omics discipline specific and data integration training course materials (Month 16)
MS3.3.2.3 Development of data integration course materials (Month 22)

Deliverables

D3.3.2.1 At least 4 different online training courses available (Month 24)
D3.3.2.1 Online training courses on all relevant topics available (Month 48)

M3.3.5 Embedding into institutional strategies

As the NFDI is focused on the whole scientific community, the embedding into the individual institutional strategies should combine general, tool specific and community oriented building blocks. Research data management should be consistent across the organisations and use as much common ground in teaching and qualification. Information hubs on RDM like forschungsdaten.org or bausteine-fdm.de could provide a good starting point for research data management commons. Research institutions should increase data literacy not only by providing special courses (compare to M3.3.1 to 3.3.3) and online training materials (see M3.3.4) provided by DataPLANT but also by qualifying future data managers to increase the pool of available qualified cross-domain personnel and relieve the general scarcity hindering modern data management in many scientific domains. Data management and analysis should develop be fostered as fields of science to increase the outcome of today's digitized research workflows as to develop novel ones. This measure focuses as well on the exchange with other consortia and the general cross-cutting topics on training and qualification. It helps the DataPLANT participants to extend and improve the existing frameworks in their hosting institutions.

Milestones

MS3.3.2.1 Analysis of institutional landscapes and best practices for data management integration (Month 18)

Deliverables

D3.3.5.1 Liaison and Harmonization of NFDI strategies with institutional data management integrations (Month 36)
D3.3.5.2 NFDI strategies integrated into main institutions (Month 60)

4.5 Task Area 4 (Project Coordination and Management)

WP 4.1 Coordination

M4.1.1 Project office

The project office is the responsible supporting body of DataPLANT for the overall coordination of DataPLANT and is the primary point of contact for members, other consortia and international coordination. It serves to control and coordinate scientific and technical progress including risk management. It further coordinates the distribution of funds and further financial aspects of the project. It provides constant support for all administrative matters of the project and regulates the communication between the individual project partners and work packages as well as the external presentation. The project office prepares the plenary meetings of the general assembly and supports the senior management as well as the technical board in their activities. It coordinates the regular working meetings of the data stewards and topic-specific committees. Together with the senior management board, it monitors the synchronization between the partial work packages. The comprehensive documentation of the results achieved and the activities carried out during the entire project are addressed as well by this measure. Regular summaries and reports are prepared with the support of the project participants involved and presented to interested consortia within the general NFDI.

Milestones

MS4.1.1.1 Setup of the boards and committees (Month 3)
MS4.1.1.2 Provide the necessary communication and coordination infrastructure (Month 6)
MS4.1.1.3 Regular updates on work package progress (Month 11, 23, 35, 47, 59)

Deliverables

D4.1.1.1 Setup of the office as first point of contact to DataPLANT (Month 4)
D4.1.1.2 Full functioning set of communication tools and support infrastructure for the DataPLANT boards and working groups (Month 3)
D4.1.1.3 Preparation and organisation of the General Assembly meetings (Month 11, 23, 35, 47, 59)
D4.1.1.4 Regular reports on the general state and progress of DataPLANT to the General Assembly (Month 12, 24, 36, 48, 60)

M4.1.2 Coordination boards

The community driven governance of DataPLANT needs supporting coordination and support infrastructure. The task areas require interaction between participants and developers which will be channelled through working groups. Further, to implement the wider DataPLANT governance and ensure sustainable user interaction and exchange between all relevant stakeholders several committees (technical, scientific and senior management boards) are set up. The coordination boards channel the continuous exchange with the wider NFDI and international initiatives. The boards coordinate the permanent update of the DataPLANT strategy and organize the change management. Further on, if the ongoing developments require direct interaction, working group leaders will dispatch delegates to standardization bodies.

Milestones
MS4.1.2.1 Technical, scientific and senior management boards are fully operable (Month 4)
Deliverables
D4.1.2.1 Regular reports to the General Assembly and wider NFDI (Month 12, 24, 36, 48, 60)

M4.1.3 Coordination with other NFDIs

Success of the NFDI strongly depends on cooperation and regular exchange between consortia as well as the general NFDI bodies. The co-speakers will work as liaison to both supported by the office. Workpackage leaders oversee the cooperation in standardization processes through the working groups (see TA1). NFDI activities will require common efforts on data management

literacy, training and qualification programmes (see TA3). The various consortia identified a list of several cross-cutting topics where NFDI either provides direct input through own workpackages or will send delegates for further elaboration and decision making. 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 and close collaboration on identifiers and data standards. Cooperating with NFDI4Chem, the exchange of basic and molecule-specific training materials between the initiatives is planned. Further, NFDI4MSE and DataPLANT will develop common basic workflows and to foster a cultural change in their research domains. The consortia FAIRmat and DataPLANT share a sample- and workflow-centric view when it comes to handling research data. Galaxy's modular tool box for data processing and computing, allowing for a flexible integration of newly added tools promises an interesting starting point for the development of comprehensive interfaces, while at the same time ensuring the necessary adaptability. Furthermore, there are common research interests on biological and soft matter. The Both consortia plan to work together in the preparation of course materials for the data stewards. The vision is, as integration is inevitable, efforts should better start to integrate right from the beginning.

Milestones

MS4.1.3.1 Presentation of a first iteration of a matrix of cross-cutting topics and involved consortia (Month 6)
MS4.1.3.2 Updates of the matrix of cross-cutting topics (Month 18, 30, 42, 54)

Deliverables

D4.1.3.1 Coordination and participation matrix for cross-cutting topics (Month 12)
D4.1.3.2 Review of the participation and definition of to be coordinated measures (Month 24, 36, 48)

M4.1.4 Internal infrastructure

As a large-scale support, development and coordination endeavour DataPLANT requires a sound project internal infrastructure ranging from communication channels, to shared resources and development tools to the infrastructure required for education and qualification. It is planned to use a cross-location project management and ticket system that integrates all participating locations in a common ticket, time and task management and uses tools that also play a role in the implementation of DataPLANT's goals (“Authentication and Authorisation Infrastructure”, repository, versioning system). The project will make use of mailing lists and (virtual) meeting rooms for all participants, the working groups and various governance bodies, a project

management tool for work package orchestration and developer support. Further tools like wikis, project calendar and websites for documentation, outreach will be provided. The data stewards need further infrastructure to manage the requests by the users and their assignments. They will operate a (virtually distributed) helpdesk through adequate electronic channels like ticket systems, chat rooms, or similar tools as well. Several infrastructure components (like mailing lists, video conferencing, project management frameworks, ...) are already available to a certain extent. DataPLANT will use preexisting resources as much as possible (M1.3.4), contributed to the project by the providers and participants. An integration with existing technical cloud and workflow infrastructure is envisioned. The orchestration of the various working environments will be organized through the project office. The operation of the higher-level development infrastructure in the context of the project is to be administered. Further, the brought in hardware and service resources need to be opened and made available in the DataPLANT context. This may require an adaptation of existing services such as bwSFS storage system, de.NBI cloud and BinAC HPC cluster. The same applies to preexisting internal and external repositories of the participants.

Milestones

MS4.1.4.1 Coordination and distribution of tasks matched to infrastructure already available (Month 2)
MS4.1.4.2 Basic set of project infrastructure made available to the partners (Month 4)
MS4.1.4.3 Requirement analysis for connectors to existing infrastructures (Month 12)

Deliverables

D4.1.4.1 Development infrastructure, staging server is fully operable (Month 6)
D4.1.4.2 Connectivity to relevant pre-existing infrastructure is enable and fully usable (Month 18)

M4.1.5 Financial operations

The significant funds provided for DataPLANT if awarded require a responsible oversight and transparent distribution within the consortium. Various stakeholders ranging from the DFG to single research institutions are involved in financial matters. The brought in federated hardware needs to be extended and updated in regular intervals. Financial operation will be supported by the project office and overseen by the project management (senior management board and co-speakers). The internal disbursement of funds is coordinated through these bodies, the annual reports are directed at the general assembly. Financial operation takes place in separate spheres with different models. Data stewards are differently financed and assigned (see M3.2.1), compared to hardware and base level service (see M4.2.1) and developers. The financing of data stewards primarily stems from DataPLANT funding (see M3.2.1) as well as for the developers

and project support functions like the office (details in M4.1.1), outreach (compare M4.2.6) and legal support (see M3.2.3). The hardware and base level services are brought in by the providers, additional sources of funding might be added through grant applications of participants or through state support like for bwSFS or federal funds like for the de.NBI cloud hardware and operation. The further developments in the financial domain will be coordinated with the respective cross-cutting topic; the development of a sound financial model for the general NFDI is to be expected during the project run time.

Milestones

MS4.1.5.1 Gathering options of viable financing models (Month 12)
MS4.1.5.2 Preparation of DataPLANT's vision for cross-cutting activities (Month 18)

Deliverables

D4.1.5.1 Coordinated vision of DataPLANT community to be presented in the general NFDI context (Month 24)

WP 4.2 Management

M4.2.1 Infrastructure federation

DataPLANT will operate a significant infrastructure as higher level services like the DataPLANT Hub, long-term data publication and workflow execution systems as well as lower level storage and compute systems. This infrastructure needs to be regularly updated and adapted to the actual requirements. The financial resources may stem from a wide range of resources including the hosting institutions, federal infrastructure like de.NBI, state sponsored systems like BinAC HPC cluster or the bwSFS storage system to user contributed resources. To develop the DataPLANT infrastructure into a "data set attractor", we will provide certain base line capacities free of charge as one of the incentives to provide data sets and metadata descriptions above a certain agreed upon bar. To allow smooth operation under raising loads (both regarding computing demand and storage capacities) the infrastructure needs to be smoothly extensible. To raise additional funds for large scale data in the hundreds of Terabytes or huge demand in CPU processing power the resources must be provided at a reasonable price. Different models to run federated infrastructures like de.NBI, the Baden-Württemberg HPC concept or further models are in operation which are to be evaluated for the applicability as a sustainable base for DataPLANT. This requires concepts for accounting and (virtual) billing for used resources which fit into the legal and financial framework of the participants research institutions. These challenges are to be coordinated in a cross-cutting fashion together with other NFDI consortia on the general NFDI level. Independent of the actual financing model the resources have to match the user's needs.

The rules of community engagement are to be defined and models for user participation to be evaluated. In some cases, it could be attractive to scale out into commercial offerings to compensate on bottlenecks in low level services.

Milestones

MS4.2.1.1 Overview on state of the relevant infrastructure for DataPLANT (Month 8)

Deliverables

D4.2.1.1 Concept for wider infrastructure federation (Month 12)
D4.2.1.2 Model for flexible infrastructure provisioning based on distributed infrastructures (Month 24)

M4.2.2 Service description, operating and business models

DataPLANT Hub and various services (see WP2.3) are addressed to a widely distributed community and need to be properly marketed and understood. Step by step, a comprehensive service description and a corresponding service catalogue for DataPLANT will be developed and disseminated to the community. The descriptions will be developed by the technical board and discussed with the users. Regular feedback will follow the intervals of the meeting of the general assembly. The service descriptions will map the necessary services for the complete data lifecycle. To achieve a sustainable catalogue of services, the development will be coordinated with the general NFDI through cross-cutting activities. This ensures that other user communities can also access specific services provided by DataPLANT. To align the NFDI-wide and services and workflow developments with the DataPLANT community, the various offerings are examined for common aspects and characteristics in the exchange at the level of the general NFDI. The services evolvement is regularly reviewed and coordinated with the technical and senior management boards.

Milestones

MS4.2.2.1 Overview on relevant services and locations (Month 12)
MS4.2.2.1 Coordination and description on consortia-spanning services (Month 18)

Deliverables

D4.2.2.1 Description of all relevant services (Month 24)
D4.2.2.2 Coordinated business and operation model for each service (Month 48)

M4.2.3 Application of operating and business models

DataPLANT needs suitable operation and business models to become sustainable in the long run (see M4.2.7). In the federated provision of services as planned for DataPLANT agreements and the compensation of efforts and costs must be considered. The general NFDI will evaluate possible and implement agreed upon models for the NFDI wide level. The involved consortia will discuss suitable models in their cross-cutting activities. The proposed models need to be discussed in continuous connection with the community and decided upon by the general assembly and senior management board. To avoid diverging models, the integration into existing service and support structures should be as seamless as possible. The participating data centres and service providers have already gained experience with certain business and operation models, for example in connection with the operation of HPC, cloud and storage services. Parameters for accounting metrics and billing models of used or reserved resources need to be developed. Proven models from previous and current projects or experiences of large consortia need to be collected and evaluated with regard to their suitability for the operation of DataPLANT. Here it has to be clarified which (future) forms of organisation are possible and necessary for this type of service provision. This will already be tested during the project by the proposed organisational structure.

Milestones

MS4.2.3.1 Survey on operation and business models of the other consortia (Month 24)

Deliverables

D4.2.3.1 Coordinated document on operation model (Month 24)
D4.2.3.2 Coordinated document on business model (financial operation) (Month 36)
D4.2.3.3 Agreed upon operation and business model for DataPLANT base level services (Month 48)

M4.2.4 Enhancements of operations

Primarily the senior management board supported by the project office oversees ongoing (administrative) developments and coordinates required change processes. There is an ongoing change towards Open Science and Open Data fostered by a rising tide of public resources. Further, the required resources for modern research workflow increase and often cannot be provided within single groups, institutes or research facilities. Supported by technological advances we observe a paradigm shirt regarding infrastructure provisioning and financing. These developments manifests in open and shared infrastructures like de.NBI, open services like Galaxy or the availability and use of virtualized research environments (10.15496/publikation-25205). This is mirrored by new models of task and responsibility distribution between users and service

providers. These developments and changes directly influence both the designated community as well as the providers in the consortium. It both allows for an organisational and structural enhancement of service provisioning and relieves research groups from non-scientific overheads like market analysis, system procurement, setup and operation. Additionally, it brings down the barriers for junior researchers and smaller groups to access relevant resources for their projects. It increases the dynamic of progress in research and the application of novel methods and workflows. Such processes need to be monitored, evaluated, agreed upon with the users and put into the enhancement of the existing business and operation models. As data stewards are a significant resource and core concept of DataPLANT, their assignment model (see M3.2.1) needs a regular review beside infrastructure operation models. The model has to manage the expectations of the stakeholders and to cope with fluctuating demand, rising awareness on RDM in the community, thus a shift of tasks and new users. The review needs to incorporate the feedback by the users.

Milestones

MS4.2.4.1 Assessment on viability on relevant public services (Month 9, 21, 33, 45, 57)

Deliverables

D4.2.4.1 Regular reports on the development of public service infrastructure and the impact on DataPLANT to the General Assembly (Month 12, 24, 36, 48, 60)

M4.2.5 Risk management

DataPLANT is tightly linked to international networks and its relevance depends on the ongoing evolvement in the field of plant research. Thus, an observation of international standards and tools developments is necessary, and a change process for the adaptation of standards is to be coordinated. The technological landscape needs to be monitored as well for ground-breaking changes for new instruments or algorithms which influence decisions and the direction of further developments. Public online repositories and services used by the DataPLANT community might disappear. The effects need to be mitigated, e.g. by migrating data sets to the DataPLANT storage or increase the level of redundant copies. Vice versa the DataPLANT services rely on a sustainable infrastructure dependent on the operating providers. Users expect the data to be stored reliably and unaltered as trust into the system and research results significantly depends on these characteristics. DataPLANT as a distributed project with numerous dependencies, spanning diverse research institutions faces risks in the organisational domain as well. Primarily, the hiring of qualified personnel in a highly competitive environment could be difficult, as well that core personnel is leaving because of the short-term contracts offered. Further, the sustainability

might become a risk if no viable long-term financing model for the services and infrastructure can be agreed upon and applied (see M4.1.4). The project management (senior management board and co-speakers) is responsible for risk management and is in constant exchange with the Technical and Scientific Boards. It reports regularly on the progress of developments. Problems with technical interfaces and technical infrastructure are analysed with the Technical Board and strategies for the appropriate reaction are coordinated. Method development is also regularly reviewed and coordinated with the scientific board.

Milestones

MS4.2.5.1 Report on technological developments and significant changes (Month 9, 21, 33, 45, 57)

Deliverables

D4.2.5.1 Regular reports on project risks (staffing, software and service development) to the General Assembly and presentation of mitigation strategies (Month 12, 24, 36, 48, 60)
D4.2.5.1 Regular reports on infrastructure, sustainability risks (operation model, hosting institution) to the General Assembly (Month 24, 48)

M4.2.6 Wider public relations

The outreach measures focuses both to the wider DataPLANT community (complementing M3.1.4) to extend the visibility and include potential future users. Furthermore, it addresses the wider public using appropriate online and offline channels. It will promote the NFDI vision in the plant community to inform about the ongoing evolvements, advertise for user commitment and interaction. DataPLANT will present itself with its objectives in the respective scientific institutions in the form of articles, newsfeeds, information events, and will provide information material in the form of text snippets on core technologies, standards evolvements to be used in local communication. A further stream is the creation of attractive templates for slides and posters for conferences and qualification events. It will support the community in creating DataPLANT and NFDI related materials, present arguments for agenda setting and forsters the exchange and communication with DFG to evolve the idea of the NFDI. The project office together with the data stewards and in coordination with TA3 will produce focused material for different target groups: research institutions, heads of research groups, PhDs and students in graduation. It will coordinate with core NFDI and other consortia to gain the attention and engagement of the wider scientific community and to support the information of the general public. Additionally, the applicant institutions and participants organize subject-specific workshops, focused to their colleagues of community to promote the application of the processes and workflows developed. Further, participants regularly present their findings on plant bioinformatics together with

associated workflows at CeBiTec, ISMB, ECCB or Gateways. In addition, the Galaxy team offers one-week workshops twice a year on modern processes, which have technologies and workflow management systems like Galaxy as a topic. Participants in DataPLANT also use their membership in the Galaxy Training Network and in GOBLET, a global bioinformatics education network, to exchange teaching materials and to bioinformatics curriculum. The comprehensive public relations activities are accompanied by attendance at field specific conferences on e.g. standardization and metadata. At professional events both in the field of information and data management as well as bioinformatics conferences, corresponding lectures will be held to teach on the achieved evolvements in DataPLANT. In the course of the project, handbooks will be prepared on the technologies and skills. and manuals that support further use of the findings. To provide momentum to the various activities of DataPLANT and efficiently disseminate developments coordinated activities will prepare and summarize events through blog posts and other appropriate media.

Milestones

MS4.2.6.1 Setup of the relevant online and offline communication channels (Month 6)
MS4.2.6.2 Overview on relevant conferences, workshops and general events (Month 8)
MS4.2.6.3 Planning of a holistic communication strategy involving the community, following all relevant events (Month 12)
MS4.2.6.4 Text snippets and information bits on ongoing developments (Month 18)

Deliverables

D4.2.6.1 Press and publication templates (Month 6)
D4.2.6.2 Preliminary information material on DataPLANT services (Month 24)
D4.2.6.3 Handbooks and further information material on DataPLANT services (Month 60)

M4.2.7 Sustainability

This measure addresses contingency measures to guarantee the permanent availability of the services. It ensures that the services are designed in such a way as to enable them to adapt to changing needs in the designated community. The sustainability of DataPLANT rests on several pillars. It depends on an ongoing relevance of the project to the designated community and a permanent user involvement and feedback. It must adapt to ongoing changes and developments in the field of plant research. The offered services require a viable financial model. The financial operations need to be updated to the suggestions produced from cross-cutting activities on governance and sustainability. Proven models of cooperation will be jointly analysed as well in the course of cross-cutting activities and evaluated with regard to their suitability for the long-term and sustainable operation of NFDI infrastructures and further developed to a common basis. In

this context, it must be agreed how the desired organisational form of the NFDI as a whole can be brought into line with DataPLANT. We work on these topics in permanent coordination with the community. For the provision of RDM services, it must be evaluated which options are available and which can be applied by the consortium as well as DataPLANT services that are also provided for other users. In particular, possible cooperation's with other existing or planned NFDI consortia dealing with RDM, cloud and HPC should be seeked. To support long-term availability of at least the base layer data storage and publication services a long-term cost compensation is required. The general NFDI and the activities in the cross-cutting topic on governance will suggest and develop governance and operation models which needs to be integrated and adapted for DataPLANT.

Milestones

MS4.2.7.1 Input to cross-cutting activities regarding financial model and sustainability to the general NFDI (Month 12)
MS4.2.7.2 Coordination with other consortia on base level compute and storage infrastructure provisioning (Month 18)
MS4.2.7.3 Input on common RDM service matrix to the general NFDI (Month 24)

Deliverables

D4.2.7.1 Coordinated document on sustainable service operation for DataPLANT (Month 30) D4.2.7.1 Application of the general NFDI sustainable service model (Month 60)

M4.2.8 Data protection, data security and certification of infrastructures

Most of the data sets relevant in DataPLANT are not affected by privacy or sensitivity considerations. Nevertheless, for federated provisioning of services, as offered within the framework of DataPLANT, aspects such as data protection and security as well as service agreements and the balancing of expenses and costs must be considered as well. There will be a continuous support for researchers (see M3.2.3) and a seamless integration into existing support structures as far as possible. Here, experience already exists both at the national level in the area of cloud infrastructures (de.NBI) and in field specific cooperation. Initially, the primary focus is on data sets from the participants. Much of the data is expected to be non-critical. Nevertheless, as data sets will get permanently linked to its creators the requirements of the EU-GDPR and the administrative regulation on information security are to be honoured. Legal support (see M3.2.2) might identify sensitive information or data sets. To properly deal with such requirements a certification of infrastructure could be helpful. The same applies for repositories and services for long-term data publication with the opportunity to cite data sets and workflows.

Here, a certification of trustworthiness might be required by journals, institutions and science funders.

Milestones

MS4.2.8.1 Report on the need for EU-GDPR, handling of sensitive data (Month 12)
MS4.2.8.1 Report on requirements on data repositories (Month 18)

Deliverables

D4.2.8.1 Outline of certification options of infrastructures regarding EU-GDPR (Month 18)
D4.2.8.2 Outline of certification options for trustworthy data repositories (Month 24)
D4.2.8.3 Acquiring basic certification (Month 42)
D4.2.8.4 Acquiring extended certification (Month 58)

(Co-) Applicant Contributions

The University of Freiburg contributes both personnel and infrastructure to DataPLANT. The Research Data Management Group, the eScience group as well as the professorship attached to the computer center will provide the equivalent of 2.5 FTE in support, infrastructure operation and

consulting. These persons were and are actively involved in various research projects fostering RDM like bwFLA[91], EMiL[92,93], ViCE[66,68] and CiTAR[69]. The computer centres of Freiburg and Tübingen already partnering for more than ten years in various research infrastructure research projects. The university supported by the state and the DFG heavily invests in storage capacity to provide a modern RDM for active and archived data through the Storage-for-Science[11] system. The bwHPC-S5 data management operates at the core of scientific large-scale computing and is active in creating a state-wide data federation. The contributed personnel and hardware resources amount to nearly 2.000.000 € over the period of five years.

The FZJ IBG-4 operates several databases and develops analysis and visualization tools in the field of plant omics data (such as Plabi.PD, CSB.DB and the MapMan Website of Tools), which are geared mostly towards genomics, transcriptomics and cross-omics analysis profiting from plant pathway data. IBG-4 will provide the necessary personnel that is needed to ensure the continuous operation of these resources, their further development and- where not yet done- their migration to the Galaxy platform. In addition, IBG-4 will install a new Plant Data Science group based on own funding whose task will be mainly in developing common standards workflows and ontologies within the phenotyping and plant bioinformatics communities in TA1. In addition, the group will contribute about 500 TB of storage and up to 256 CPU cores. In sum this contribution amounts to ~900.000 € over five years.

The Backofen Group contributes ~1000 CPU cores to the DataPLANT project from own funding. Furthermore, Prof. Rolf Backofen will tighten the connections to ELIXIR as ELIXIR Germany Board Member. In sum, this contribution amounts to ~750.000 €.

The computing centre of the University of Kaiserslautern contributes to the project the basic data storage and data processing infrastructure - specifically a dedicated virtualization platform in combination with adequate high-performance compute capabilities as well as ~400 TB of redundant hard disk storage and tape backup for short and long-term data storage. In sum, this contribution amounts to ~500.000 €.

The computing centre of the University of Tübingen contributes compute and storage resources to the DataPLANT project as well as operating personnel. Up to 250 cores and 2 PB of storage may be used from the de.NBI Cloud. For computing purposes up to 850 cores of the BinAC cluster are available. The successor system BinAC II can be used in a similar way upon availability. Storage space of up to 2 PB may be occupied on the Storage-for-Science which is operated jointly with the University of Freiburg. In sum, the overall contribution amounts to ~1.7960.000 €.

6 General Compliance

We ensure that the information provided through this proposal submitted to the Deutsche Forschungsgemeinschaft is accurate.

The applicant/co-applicant institutions confirm that all persons participating in the consortium from this institution agree to adhere to the DFG's rules of good scientific practice. The DFG's Rules of Procedure for Dealing with Scientific Misconduct (Verfahrensordnung zum Umgang mit wissenschaftlichem Fehlverhalten – VerfOwF) apply to individuals with a high level of scientific responsibility in funding proposals submitted to the DFG by higher education institutions and non-university institutions. In signing this compliance form, the applicant/co-applicant institution and the spokesperson and co-spokespersons of the proposed consortium acknowledge and recognize as legally binding the aforementioned DFG Rules of Procedure. If, during the course of the funding period, the spokesperson or co-spokesperson changes, please note that the new person must sign a declaration of obligation of compliance, which must be forwarded to the DFG upon request in accordance with its Rules of Procedure. The applicant/co-applicant institution agrees to the DFG's electronic processing and storage of data provided in conjunction with this proposal. It further agrees to this information being used for evaluation and statistical purposes and forwarded to reviewers and decision-making bodies – which may involve a third country – as part of the review and decision-making process.

The applicant/co-applicant institution ensures that all individuals listed in the proposal (including participants in the case of the applicant institution) agree to this and to the forwarding of the final funding decisions to the head of the institution and the (co-)spokespersons.

We agree that, if the proposal is approved, our work addresses and contact details (name, institution and location, phone, fax, e-mail and website) as well as information about the content of the project (e.g., topic, summary, keywords, subject area, program, funding period, international connections) will be published in the GEPRIS information system and may be published in other, non-commercial publications and databases created in cooperation with the DFG. We understand that we may withdraw our consent to full/partial publication at any time without affecting the lawfulness of any processing carried out prior to our withdrawal by notifying the responsible DFG contact, preferably in electronic form.

7 Appendix

The appendix may only include the following information and documents:

1 Bibliography and list of references

1. FAIR principles for data stewardship. *Nature genetics* **48,** 343; 10.1038/ng.3544 (2016).

2. Digital Science *et al.* The State of Open Data Report 2018.

3. Afgan, E. *et al.* The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic acids research* **46,** W537-W544; 10.1093/nar/gky379 (2018).

4. Tauch, A. & Al-Dilaimi, A. Bioinformatics in Germany: toward a national-level infrastructure. *Briefings in bioinformatics* **20,** 370–374; 10.1093/bib/bbx040 (2019).

5. Lynch, C. Big data: How do your data grow? *Nature* **455,** 28–29; 10.1038/455028a (2008).

6. Borgman, C. L. The conundrum of sharing research data. *Acta Anaesthesiol Scand* **63,** 1059–1078; 10.1002/asi.22634 (2012).

7. Koltay, T. Data literacy for researchers and data librarians. *Journal of Librarianship and Information Science* **49,** 3–14; 10.1177/0961000615616450 (2017).

8. Leonelli, S., Davey, R. P., Arnaud, E., Parry, G. & Bastow, R. Data management and best practice for plant science. *Nature plants* **3,** 17086; 10.1038/nplants.2017.86 (2017).

9. Langmead, B. & Nellore, A. Cloud computing for genomic data analysis and collaboration. *Nature reviews. Genetics* **19,** 208–219; 10.1038/nrg.2017.113 (2018).

10. Belmann, P. *et al.* de.NBI Cloud federation through ELIXIR AAI. *F1000Research* **8,** 842; 10.12688/f1000research.19013.1 (2019).

11. Suchodoletz, D. v., Hahn, U., Wiebelt, B., Glogowski, K. & Seifert, M. Storage infrastructures to support advanced scientific workflows. Towards research data management aware storage infrastructures.

12. Korte, A. & Farlow, A. The advantages and limitations of trait analysis with GWAS: a review. *Plant methods* **9,** 29; 10.1186/1746-4811-9-29 (2013).

13. Schwacke, R. *et al.* MapMan4: A Refined Protein Classification and Annotation Framework Applicable to Multi-Omics Data Analysis. *Molecular plant* **12,** 879–892; 10.1016/j.molp.2019.01.003 (2019).

14. Grüning, B. *et al.* Bioconda: sustainable and comprehensive software distribution for the life sciences. *Nature methods* **15,** 475–476; 10.1038/s41592-018-0046-7 (2018).

15. Grüning, B. *et al.* Practical Computational Reproducibility in the Life Sciences. *Cell systems* **6,** 631–635; 10.1016/j.cels.2018.03.014 (2018).

16. Ramírez, F. *et al.* deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic acids research* **44,** W160-5; 10.1093/nar/gkw257 (2016).

17. Wolff, J. *et al.* Galaxy HiCExplorer: a web server for reproducible Hi-C data analysis, quality control and visualization. *Nucleic acids research* **46,** W11-W16; 10.1093/nar/gky504 (2018).

18. Rettberg, S., Suchodoletz, D. v. & Bauer, J. Feeding the Masses: DNBD3. Simple, efficient, redundant block device for large scale HPC, Cloud and PC pool installations.

19. Bauer, J. *et al.* A Sorting Hat For Clusters. Dynamic Provisioning of Compute Nodes for Colocated Large Scale Computational Research Infrastructures.

20. Suchodoletz, D. v., Schulz, J. C., Leendertse, J., Hotzel, H. & Wimmer, M. (eds.). *Kooperation von Rechenzentren* (De Gruyter, Berlin, Boston, 2016).

21. Münchenberg, J., Suchodoletz, D. v., Rettberg, S., Richter, S. & Rößler, C. in *Kooperation von Rechenzentren,* edited by D. v. Suchodoletz, J. C. Schulz, J. Leendertse, H. Hotzel & M. Wimmer (De Gruyter, Berlin, Boston, 2016).

22. Bauer, J. *et al. in EDULEARN19 Proceedings,* edited by L. Gómez Chova, A. López Martínez & I. Candel Torres (IATED2019), pp. 5548–5555.

23. Janczyk, M., Suchodoletz, D. v. & Wiebelt, B. bwForCluster NEMO. Forschungscluster für die Wissenschaft.

24. Batut, B. *et al.* Community-Driven Data Analysis Training for Biology. *Cell systems* **6,** 752-758.e1; 10.1016/j.cels.2018.05.012 (2018).

25. Glöckner, F. O. *et al. Berlin Declaration on NFDI Cross-Cutting Topics* (Zenodo, 2019), https://doi.org/10.5281/zenodo.3457213 Add to Citavi project by DOI.

26. Wesner, S., Walter, T., Wiebelt, B. & Suchodoletz, D. v. in *Kooperation von Rechenzentren,* edited by D. v. Suchodoletz, J. C. Schulz, J. Leendertse, H. Hotzel & M. Wimmer (De Gruyter, Berlin, Boston, 2016).

27. Wiebelt, B. *et al. in Kooperation von Rechenzentren,* edited by D. v. Suchodoletz, J. C. Schulz, J. Leendertse, H. Hotzel & M. Wimmer (De Gruyter, Berlin, Boston, 2016).

28. Linden, M. *et al.* Common ELIXIR Service for Researcher Authentication and Authorisation. *F1000Research* **7**; 10.12688/f1000research.15161.1 (2018).

29. Paskin, N. Digital object identifiers. *ISU* **22,** 97–112; 10.3233/ISU-2002-222-309 (2002).

30. Senapathi, T., Bray, S., Barnett, C. B., Grüning, B. & Naidoo, K. J. Biomolecular Reaction and Interaction Dynamics Global Environment (BRIDGE). *Bioinformatics (Oxford, England)* **35,** 3508–3509; 10.1093/bioinformatics/btz107 (2019).

31. van der Spoel, D. *et al.* GROMACS: fast, flexible, and free. *Journal of computational chemistry* **26,** 1701–1718; 10.1002/jcc.20291 (2005).

32. Cook, C. E., Bergman, M. T., Cochrane, G., Apweiler, R. & Birney, E. The European Bioinformatics Institute in 2017: data coordination and integration. *Nucleic acids research* **46,** D21-D29; 10.1093/nar/gkx1154 (2018).

33. da Veiga Leprevost, F. *et al.* BioContainers: an open-source and community-driven framework for software standardization. *Bioinformatics (Oxford, England)* **33,** 2580–2582; 10.1093/bioinformatics/btx192 (2017).

34. Hemme, D. *et al.* Systems-Wide Analysis of Acclimation Responses to Long-Term Heat Stress and Recovery in the Photosynthetic Model Organism Chlamydomonas reinhardtii. *The Plant cell* **26,** 4270–4297; 10.1105/tpc.114.130997 (2014).

35. Mettler, T. *et al.* Systems Analysis of the Response of Photosynthesis, Metabolism, and Growth to an Increase in Irradiance in the Photosynthetic Model Organism Chlamydomonas reinhardtii. *The Plant cell* **26,** 2310–2350; 10.1105/tpc.114.124537 (2014).

36. Schmollinger, S. *et al.* Nitrogen-Sparing Mechanisms in Chlamydomonas Affect the Transcriptome, the Proteome, and Photosynthetic Metabolism. *The Plant cell* **26,** 1410–1435; 10.1105/tpc.113.122523 (2014).

37. Magrane, M. & Consortium, U. UniProt Knowledgebase. a hub of integrated protein data. *Database : the journal of biological databases and curation* **2011,** bar009; 10.1093/database/bar009 (2011).

38. Bolser, D., Staines, D. M., Pritchard, E. & Kersey, P. Ensembl Plants: Integrating Tools for Visualizing, Mining, and Analyzing Plant Genomics Data. *Methods in molecular biology (Clifton, N.J.)* **1374,** 115–140; 10.1007/978-1-4939-3167-5_6 (2016).

39. van Bel, M. *et al.* PLAZA 4.0: an integrative resource for functional, evolutionary and comparative plant genomics. *Nucleic acids research* **46,** D1190-D1196; 10.1093/nar/gkx1002 (2018).

40. Butler, D. Gates Foundation announces open-access publishing venture. *Nature* **543,** 599; 10.1038/nature.2017.21700 (2017).

41. Krishnakumar, V. *et al.* Araport: the Arabidopsis information portal. *Nucleic acids research* **43,** D1003-9; 10.1093/nar/gku1200 (2015).

42. Schwacke, R. *et al.* ARAMEMNON, a novel database for Arabidopsis integral membrane proteins. *Plant physiology* **131,** 16–26; 10.1104/pp.011577 (2003).

43. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature genetics* **25,** 25–29; 10.1038/75556 (2000).

44. Usadel, B. *et al.* A guide to using MapMan to visualize and compare Omics data in plants: a case study in the crop species, Maize. *Plant, cell & environment* **32,** 1211–1229; 10.1111/j.1365-3040.2009.01978.x (2009).

45. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic acids research* **28,** 27–30; 10.1093/nar/28.1.27 (2000).

46. Afendi, F. M. *et al.* KNApSAcK family databases: integrated metabolite-plant species databases for multifaceted plant research. *Plant & cell physiology* **53,** e1; 10.1093/pcp/pcr165 (2012).

47. Heazlewood, J. L., Verboom, R. E., Tonti-Filippini, J., Small, I. & Millar, A. H. SUBA: the Arabidopsis Subcellular Database. *Nucleic acids research* **35,** D213-8; 10.1093/nar/gkl863 (2007).

48. Hettne, K. M. *et al.* Structuring research methods and data with the research object model: genomics workflows as a case study. *Journal of biomedical semantics* **5,** 41; 10.1186/2041-1480-5-41 (2014).

49. Eoghan Ó Carragáin, Carole Goble, Peter Sefton & Stian Soiland-Reyes. A lightweight approach to research object data packaging.

50. González-Beltrán, A., Maguire, E., Sansone, S.-A. & Rocca-Serra, P. linkedISA: semantic representation of ISA-Tab experimental metadata. *BMC bioinformatics* **15 Suppl 14,** S4; 10.1186/1471-2105-15-S14-S4 (2014).

51. Vizcaíno, J. A. *et al.* ProteomeXchange provides globally co-ordinated proteomics data submission and dissemination. *Nature biotechnology* **32,** 223–226; 10.1038/nbt.2839 (2014).

52. Clough, E. & Barrett, T. The Gene Expression Omnibus database. *Methods in molecular biology (Clifton, N.J.)* **1418,** 93–110; 10.1007/978-1-4939-3578-9_5 (2016).

53. Leinonen, R., Sugawara, H. & Shumway, M. The sequence read archive. *Nucleic acids research* **39,** D19-21; 10.1093/nar/gkq1019 (2011).

54. Haug, K. *et al.* MetaboLights—an open-access general-purpose repository for metabolomics studies and associated meta-data. *Nucleic acids research* **41,** D781-6; 10.1093/nar/gks1004 (2012).

55. re3data.org. figshare.

56. Castro, E. Dataverse is now minting DOIs with DataCite Metadata Store API.

57. Taylor, C. F. *et al.* Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the MIBBI project. *Nature biotechnology* **26,** 889–896; 10.1038/nbt.1411 (2008).

58. Zimmermann, P. *et al.* MIAME/Plant - adding value to plant microarrray experiments. *Plant methods* **2,** 1; 10.1186/1746-4811-2-1 (2006).

59. Ćwiek-Kupczyńska, H. *et al.* Measures for interoperability of phenotypic data: minimum information requirements and formatting. *Plant methods* **12,** 44; 10.1186/s13007-016-0144-4 (2016).

60. Dzale Yeumo, E. *et al.* Developing data interoperability using standards: A wheat community use case. *F1000Research* **6,** 1843; 10.12688/f1000research.12234.2 (2017).

61. Peter Amstutz *et al.* Common Workflow Language, v1.0.

62. Brazma, A. *et al.* Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nature genetics* **29,** 365–371; 10.1038/ng1201-365 (2001).

63. Field, D. *et al.* The minimum information about a genome sequence (MIGS) specification. *Nature biotechnology* **26,** 541–547; 10.1038/nbt1360 (2008).

64. Fiehn, O. *et al.* Minimum reporting standards for plant biology context information in metabolomic studies. *Metabolomics* **3,** 195–201; 10.1007/s11306-007-0068-0 (2007).

65. Martínez-Bartolomé, S., Binz, P.-A. & Albar, J. P. The Minimal Information about a Proteomics Experiment (MIAPE) from the Proteomics Standards Initiative. *Methods in molecular biology (Clifton, N.J.)* **1072,** 765–780; 10.1007/978-1-62703-631-3_53 (2014).

66. Perez-Riverol, Y. *et al.* Quantifying the impact of public omics data. *Nature communications* **10,** 3512; 10.1038/s41467-019-11461-w (2019).

67. Piwowar, H. A., Day, R. S. & Fridsma, D. B. Sharing detailed research data is associated with increased citation rate. *PloS one* **2,** e308; 10.1371/journal.pone.0000308 (2007).

68. Bartusch, F., Hanussek, M. & Krüger, J. Containerization of Galaxy Workflows increases Reproducibility.

69. Belhajjame, K. *et al.* (eds.). *Workflow-centric research objects: First class citizens in scholarly discourse* (2012).

70. Kurtzer, G. M., Sochat, V. & Bauer, M. W. Singularity: Scientific containers for mobility of compute. *PloS one* **12,** e0177459; 10.1371/journal.pone.0177459 (2017).

71. Janczyk, M., Wiebelt, B. & Suchodoletz, D. v. Virtualized Research Environments on the bwForCluster NEMO.

72. Bührer, F. *et al.* Dynamic Virtualized Deployment of Particle Physics Environments on a High Performance Computing Cluster. *Comput Softw Big Sci* **3,** 92056; 10.1007/s41781-019-0024-5 (2019).

73. Bauer, J., Suchodoletz, D. v., Vollmer, J. & Rasche, H. Game of Templates. Deploying and (re-)using Virtualized Research Environments in High-Performance and High-Throughput Computing.

74. Wehrle, D. & Rechert, K. Are Research Datasets FAIR in the Long Run? *IJDC* **13,** 294–305; 10.2218/ijdc.v13i1.659 (1970).

75. Nachrichten. *ABI Technik* **36**; 10.1515/abitech-2016-0040 (2016).

76. Ison, J. *et al.* EDAM: an ontology of bioinformatics operations, types of data and identifiers, topics and formats. *Bioinformatics (Oxford, England)* **29,** 1325–1332; 10.1093/bioinformatics/btt113 (2013).

77. Hastings, J. *et al.* ChEBI in 2016: Improved services and an expanding collection of metabolites. *Nucleic acids research* **44,** D1214-9; 10.1093/nar/gkv1031 (2016).

78. Walls, R. L. *et al.* The Plant Ontology Facilitates Comparisons of Plant Development Stages Across Species. *Frontiers in plant science* **10,** 631; 10.3389/fpls.2019.00631 (2019).

79. FAIRsharing Team. Plant Trait Ontology.

80. Vizcaíno, J. A. *et al.* 2016 update of the PRIDE database and its related tools. *Nucleic acids research* **44,** D447-56; 10.1093/nar/gkv1145 (2016).

81. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics (Oxford, England)* **26,** 139–140; 10.1093/bioinformatics/btp616 (2010).

82. Shah, N., Nute, M. G., Warnow, T. & Pop, M. Misunderstood parameter of NCBI BLAST impacts the correctness of bioinformatics workflows. *Bioinformatics (Oxford, England)* **35,** 1613–1614; 10.1093/bioinformatics/bty833 (2019).

83. Harrison, P. W. *et al.* The European Nucleotide Archive in 2018. *Nucleic acids research* **47,** D84-D88; 10.1093/nar/gky1078 (2019).

84. Arend, D. *et al.* PGP repository: a plant phenomics and genomics data publication infrastructure. *Database : the journal of biological databases and curation* **2016**; 10.1093/database/baw033 (2016).

85. Di Tommaso, P. *et al.* Nextflow enables reproducible computational workflows. *Nature biotechnology* **35,** 316–319; 10.1038/nbt.3820 (2017).

86. Solbrig, H. R. *et al.* Modeling and validating HL7 FHIR profiles using semantic web Shape Expressions (ShEx). *Journal of biomedical informatics* **67,** 90–100; 10.1016/j.jbi.2017.02.009 (2017).

87. Cruz, A., Arrais, J. P. & Machado, P. Interactive and coordinated visualization approaches for biological data analysis. *Briefings in bioinformatics* **20,** 1513–1523; 10.1093/bib/bby019 (2019).

88. Kerren, A., Kucher, K., Li, Y.-F. & Schreiber, F. MDS-based Visual Survey of Biological Data Visualization Techniques.

89. Kerren, A., Kucher, K., Li, Y.-F. & Schreiber, F. BioVis Explorer: A visual guide for biological data visualization techniques. *PloS one* **12,** e0187341; 10.1371/journal.pone.0187341 (2017).

90. Crisan, A., Gardy, J. L. & Munzner, T. A systematic method for surveying data visualizations and a resulting genomic epidemiology visualization typology: GEViT. *Bioinformatics (Oxford, England)* **35,** 1668–1676; 10.1093/bioinformatics/bty832 (2019).

91. Callahan, S. P. *et al. in ACM SIGMOD/PODS 2006,* edited by C. Yu, P. Scheuermann & S. Chaudhuri (Association for Computing Machinery, [New York], 2006), p. 745.

92. North, C. *et al. in Proceedings of the 2011 annual conference extended abstracts on Human factors in computing systems - CHI EA '11,* edited by D. Tan, S. Amershi, B. Begole, W. A. Kellogg & M. Tungare (ACM Press, New York, New York, USA, 2011), p. 33.

93. Davidson, S. B. & Freire, J. in *SIGMOD-PODS '08,* edited by J. Wang (ACM, New York, NY, 2009), p. 1345.

94. Ragan, E. D., Endert, A., Sanyal, J. & Chen, J. Characterizing Provenance in Visualization and Data Analysis: An Organizational Framework of Provenance Types and Purposes. *IEEE transactions on visualization and computer graphics* **22,** 31–40; 10.1109/TVCG.2015.2467551 (2016).

95. *2019 IEEE/ACM 16th International Conference on Mining Software Repositories (MSR)* (IEEE, 2019 - 2019).

96. Rechert, K., Valizada, I., Suchodoletz, D. v. & Latocha, J. bwFLA – A Functional Approach to Digital Preservation. *PIK - Praxis der Informationsverarbeitung und Kommunikation* **35**; 10.1515/pik-2012-0044 (2012).

97. Rechert, K., Liebetraut, T., Stobbe, O., Lubetzki, N. & Steinke, T. The RESTful EMiL. *Alexandria* **27,** 120–136; 10.1177/0955749017725427 (2017).

98. Rechert, K. *et al.* Take care of your belongings today – securing accessibility to complex electronic business processes. *Electron Markets* **62,** 1009; 10.1007/s12525-013-0151-5 (2014).