the plant journal
SEB SOCIETY FOR EXPERIMENTAL BIOLOGY

The Plant Journal (2023) **116,** 974–988 doi: 10.1111/tpj.16474

RESOURCE

PLANTdataHUB: a collaborative platform for continuous FAIR data sharing in plant research

Heinrich Lukas Weil[1,†], Kevin Schneider[1,†], Marcel Tschöpe[2], Jonathan Bauer[2], Oliver Maus[1], Kevin Frey[1], Dominik Brilhaus[3], Cristina Martins Rodrigues[2], Gajendra Doniparthi[4], Florian Wetzels[5], Jonas Lukasczyk[5], Angela Kranz[6], Björn Grüning[7], David Zimmer[1], Stefan Deßloch[4], Dirk von Suchodoletz[2], Björn Usadel[3,6], Christoph Garth[5] and Timo Mühlhaus[1,*]

[1]*Computational Systems Biology, University of Kaiserslautern-Landau, Kaiserslautern, Germany,*
[2]*Computer Center, University of Freiburg, Freiburg im Breisgau, Germany,*
[3]*Cluster of Excellence on Plant Sciences (CEPLAS), Heinrich Heine University Düsseldorf, Düsseldorf, Germany,*
[4]*Heterogenous Information Systems, University of Kaiserslautern-Landau, Kaiserslautern, Germany,*
[5]*Scientific Visualization Lab, University of Kaiserslautern-Landau, Kaiserslautern, Germany,*
[6]*IBG-4 Bioinformatics, BioSC, Forschungszentrum Jülich, Jülich, Germany, and*
[7]*Bioinformatics Group, University of Freiburg, Freiburg im Breisgau, Germany*

Received 2 September 2023; accepted 8 September 2023; published online 11 October 2023.
*For correspondence (e-mail timo.muehlhaus@rptu.de).
[†]These authors contributed equally.

SUMMARY

In modern reproducible, hypothesis-driven plant research, scientists are increasingly relying on research data management (RDM) services and infrastructures to streamline the processes of collecting, processing, sharing, and archiving research data. FAIR (i.e., findable, accessible, interoperable, and reusable) research data play a pivotal role in enabling the integration of interdisciplinary knowledge and facilitating the comparison and synthesis of a wide range of analytical findings. The PLANTdataHUB offers a solution that realizes RDM of scientific (meta)data as evolving collections of files in a directory – yielding FAIR digital objects called ARCs – with tools that enable scientists to plan, communicate, collaborate, publish, and reuse data on the same platform while gaining continuous quality control insights. The centralized platform is scalable from personal use to global communities and provides advanced federation capabilities for institutions that prefer to host their own satellite instances. This approach borrows many concepts from software development and adapts them to fit the challenges of the field of modern plant science undergoing digital transformation. The PLANTdataHUB supports researchers in each stage of a scientific project with adaptable continuous quality control insights, from the early planning phase to data publication. The central live instance of PLANTdataHUB is accessible at (https://git.nfdi4plants.org), and it will continue to evolve as a community-driven and dynamic resource that serves the needs of contemporary plant science.

Keywords: ARC Annotated research context, DataHUB, FAIR, FAIR digital object, research data management.

INTRODUCTION

Scientific data is the foundation and simultaneously primary product of modern research and enables researchers to test hypotheses, provide evidence, identify patterns and trends, replicate experiments, and support innovation. The increasingly interdisciplinary landscape of science requires an ever-expanding skill set. While research data management (RDM) has been rightfully identified as one of the key areas for improving reproducibility and accessibility in the digital age, it must be ensured that solutions to emerging challenges are not only technical solutions but fit the reality of how research is usually conducted in the field of plant research.

Within this context, FAIR data (Mayer et al., 2021; Wilkinson et al., 2016) (i.e., data that is findable, accessible, interoperable, and reusable) helps to ensure that research data is managed in a way that maximizes its value and enables broader transparency, and especially reproducibility and collaboration. Modern research in plant science is dependent upon the ability to access and integrate (heterogeneous) data from published sources as well as ongoing experimental investigations. Besides the obvious raw expression data databases NCBI and EBI, plant data from multiple sources has been successfully integrated into, for example, the eFP plant browser which allows the visualization of published data (Winter et al., 2007), CoNekT, which allows transcriptomic network building (Proost & Mutwil, 2018) whereas PlabiPd lists all available Plant Genomes (Schwacke et al., 2019). Further, technical advances paired with the inherent biological complexity suggest collaborative, and interdisciplinary research as a future-oriented success model.

Currently, there are many public repositories to store biological data of a particular type, for example transcriptomics (Barrett & Edgar, 2006; Cochrane et al., 2016; Parkinson et al., 2007), proteomics (Hermjakob & Apweiler, 2006; Vizcaíno et al., 2010), metabolomics (Haug et al., 2017, 2020), or knowledge about a specific model organism (Krishnakumar et al., 2015; Lawrence et al., 2004). However, this domain and discipline storage of data is not conducive for modern investigations as interdependencies between multiple measurement techniques, experimental protocols, and computational workflows are essential, especially in cases where multiple omics datasets [e.g., metabolomics and transcriptomics, see (Zhu et al., 2023)] were generated from the same samples.

Especially in recent plant science, a huge variety of methods are used to address-specific research questions and technologies are employed driven by the research question. A FAIR data repository for plant research needs to be technology-agnostic and multimodal in which data is organized according to a research question and not separated due to measurement technology or data type generated. Further, a single uniform data annotation that does not depend on the technology-specific target repository is required to minimize the data documentation workload on the site of the plant researcher.

Continuous evolution and collaboration in scientific inquiry

Scientific research thrives on the dynamic and ever-evolving nature of information. While raw measurement data remains immutable by default, the research context continually expands through experimentation and collaboration with peers. This dynamic landscape necessitates a paradigm shift in the perception of research data, from immutable artifacts to constantly evolving entities. Continuous evaluation and analysis are paramount in this context, ensuring the relevance and accuracy of research outcomes. To facilitate this, the technical capability for continuous data annotation is imperative. Such annotation serves to keep data up-to-date and aligned with the evolving research narrative.

Data curation emerges as a critical task, requiring expertise and potentially leveraging crowdsourcing (Zhou et al., 2018) to effectively manage and maintain research data. Open-source collaboration further enhances the collective endeavor, allowing researchers to contribute, share, and harness the full potential of evolving datasets. Therefore, the dynamic nature of research data emphasizes the need for ongoing evaluation, technical adaptability, expert curation, and collaborative efforts in the pursuit of scientific knowledge.

Additionally, the collaborative nature of scientific research demands effective project management tools to harness the collective potential of research teams. In this context, a software system designed for project management becomes invaluable to offer features for task management with issue tracking, and transparent communication, all of which are essential for collaborative endeavors in science. Task management allows researchers to define and assign workloads to ensure carefully planned and delegated contributions, especially in interdisciplinary projects. Furthermore, the possibility of systematically tracking issues provides a central repository for documenting problems, their solutions, and lessons learned. This not only aids in resolving current issues but also serves as a knowledge base for future reference. Effective and transparent communication renders the cornerstone of successful collaboration and should be an integral part of a modern FAIR RDM platform.

Data-centric RDM: balancing flexibility and accessibility in a global scientific landscape

Plant research is becoming increasingly data-centric, relying on seamless integration and access to diverse and extensive datasets (Krantz et al., 2021). However, the dynamic and ever-evolving nature of research data poses significant challenges in achieving a harmonious equilibrium between data-centricity and global accessibility, with a particular emphasis on a flexible data infrastructure. Acknowledging the impossibility of anticipating all feature requirements, a flexible foundation is essential. The focus should be on building a data-centered ecosystem, with a data pool that minimizes technical overhead. This approach enables data to be imported and consumed by specific databases and services without cumbersome adaptations. Access to data should be facilitated through Application Programming Interfaces (APIs) and software solutions that manage complete research artifacts. This allows for direct data access, empowering researchers to integrate their tools seamlessly and foster collaboration. Also, ensuring compatibility with other established domain-based repositories is vital. Especially technology-specific repositories like ProteomeExchange (Hermjakob & Apweiler, 2006), MetaboLights (Haug et al., 2020), or INSDC (Cochrane et al., 2016) should be integrated into a multimodal approach to leverage established infrastructures and knowledge clusters. However, to free researchers from the onerous task of mastering not one but multiple data annotation techniques (e.g., one for transcriptomics data and a different one for proteomics data) and adhering to diverse platform-specific guidelines, it is imperative to unify data annotation logic. While tools and APIs seem an easy route to standardization of RDM, researchers must retain sovereignty over their data, avoiding technological lock-in and safeguarding data.

Integrated infrastructure for collaborative RDM

RDM is frequently perceived as an additional burden, requiring a substantial investment of time on top of the typical research workload. An effective strategy for integrating RDM into a researcher's workflow is through collaboration on research data, whereby the research process drives data management. In the context of modern plant research, which thrives on collaborative efforts, data sharing with immediate collaborators becomes pivotal. Sharing structured data with direct access facilitates not only the dissemination of information to collaborators but also aids in the later steps of achieving FAIR data and public data sharing.

While researchers frequently utilize public repositories, data management typically follows a 'local-first' approach, as data originates within individual research institutions. Decentralized data storage not only enhances data access speed due to physical proximity but also aligns with the principle of local data generation and analysis. However, for the purpose of publication, data centralization in data hubs is essential, which currently results in a disjointed process with redundant efforts. To address this, local infrastructure should closely resemble the central infrastructure in look and feel, ensuring a cohesive user experience. Moreover, local infrastructure should be capable of integrating seamlessly with the same API utilized in the central infrastructure. Additionally, data findability must be facilitated through well-established data harvesters such as search engines [e.g., Google Dataset Search (Brickley et al., 2019), the FAIDARE data discovery portal (Pommier et al., 2023), or OpenAIRE (Rettberg & Schmidt, 2012)], streamlining the process of discovering and accessing research data. Efficient sharing across institutional boundaries necessitates the ability to grant access rights seamlessly. The complexities surrounding data access rights mandate meticulous attention. The establishment and implementation of a robust and most importantly user-friendly Authentication and Authorization Infrastructure (AAI) allowing to use existing logins has become indispensable for effectively managing and enforcing data access policies.

Ensuring data quality in collaborative RDM

In collaborative research projects involving multiple users working on a shared project, effective version control and mechanisms for proposing and incorporating changes to a project are pivotal. These mechanisms should strike a balance between granting users the capability to suggest modifications, while providing principal investigators with control to evaluate and influence the quality of these proposed changes. Data quality within such collaborative research projects can be dissected into two primary facets: Firstly, (meta)data completeness and methodological or even experimental data quality. Completeness of metadata and data documentation can be assessed by automated processes thoroughly scanning (meta)data and reporting back to researchers on the status of data completeness, streamlining this critical aspect of data quality management.

Conversely, the second facet, encompassing methodical and experimental data quality, typically necessitates some degree of project management oversight and expert user interaction. However, this user-driven element can also be substantially bolstered by automated processes. Here, automation can support recurring tasks and standard data analysis workflows that are contingent upon the specific data and measurement involved in the research. For instance, quality assessments such as evaluating the replicability of results or gauging genome/proteome coverage could be integrated into these automated processes, thereby enhancing data quality assurance within the collaborative research framework.

FAIRification and publication of research data

The requirements and foundation of a successful RDM platform in the field of plant science can be distilled from the FAIR data principles, effectively met through the generation of so-called FAIR digital objects (FDOs) on a collaborative platform. FDOs establish a model for representing digital entities describing data and their metadata together with mechanisms for their creation, maintenance, and reuse, in accordance with FAIR principles. Thus, FDOs consolidate all critical information about a given entity into a unified, technology-independent object, enriching our digital landscape.

Nonetheless, the journey towards FAIRification of research data, a process easily achievable by creating FDOs, is a progressive one that advances from raw data towards a comprehensive contextualized data publication. This necessitates an RDM platform capable of facilitating this transformation. An effective RDM platform aspires to empower users by converting their research data into FAIR digital objects within a seamlessly interconnected environment. Such a platform essentially curates a resource that amalgamates experimental outcomes into vast knowledge networks, elevating data to the status of first-class contributions to scientific knowledge dissemination and transfer. This transition is not only significant for extensive omics experiments but also for the meticulous integration of findings from smaller-scale experiments, culminating in substantial knowledge accumulation.

In combination with journal publications that reference FDOs, data publications possess distinct advantages, notably in terms of ease of reuse and the creation of highly valuable research entities. These advantages are particularly pronounced in the context of contemporary machine-learning techniques and AI-driven knowledge discovery. Further, FDO publications offer numerous advantages to the scientific community. They enhance visibility and citations, accelerate the publication process, foster collaboration, make negative results publishable, and reduce redundant experimental efforts. By embracing FDOs, researchers contribute to a more efficient, transparent, and collaborative research ecosystem, ultimately advancing the frontiers of scientific knowledge.

PLANTdataHUB - a fully featured, end-to-end solution to FAIR RDM in plant research

PLANTdataHUB is a platform for working with research data organized in ARCs (*Annotated Research Contexts*), which represent biology-targeted Fair Digital Objects (DataPLANT Community, 2023a). By combining well-proven open standards, ARCs provide a data-centric approach with low overhead for researchers. They are methodology-agnostic and can support data in any format, measurements from any tool, and computation in any environment. ARCs fundamentally enable collaboration, crowdsourcing, and a continuous perspective on data curation, while still retaining effortless interoperability with standard data repositories.

The PLANTdataHUB platform provides a resource, enabling modern RDM for the plant community with state-of-the art knowledge organization. It combines existing free and open-source platforms and standards to store and organizes ARCs in a central repository. Moreover, it supports ARCs with facilities for automated quality control, data interrogation, and collaboration workflows. PLANTdataHUB is also a software-as-a-service blueprint, and instances can be realized at any level (personal, institutional, or global) that is convenient for the supported research process, while retaining full interoperability with the central instance and each other. In this manuscript, we describe the architecture of the PLANTdataHUB and explore the design choices underlying it.

RESULTS

We introduce PLANTdataHUB as a comprehensive solution tailored for the plant research community, implementing the Hub blueprint. At its core, PLANTdataHUB comprises four key components: a customized GitLab instance, a central Keycloak identity provider, InvenioRDM, and an ARC registry for ARC management and searching (Figure 1). The central instance described here is readily available, providing a full suite of functionalities, and is actively maintained by the DataPLANT NFDI4plants initiative. This primary instance seamlessly integrates with InvenioRDM for data publication, which can be initiated directly from PLANTdataHUB. For collaborative research, ARCs can be configured as private, restricted to selected collaborators,

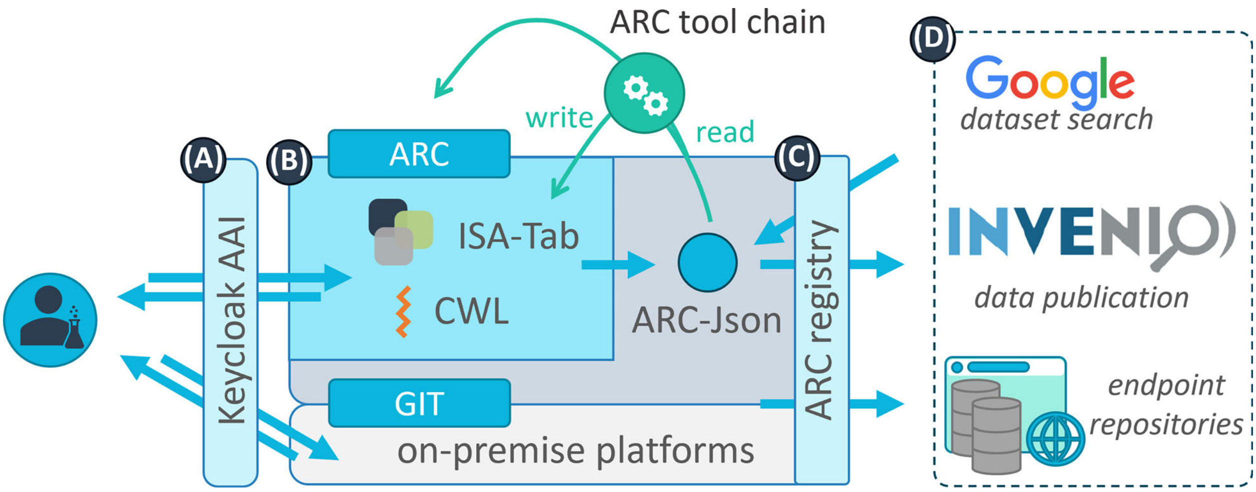
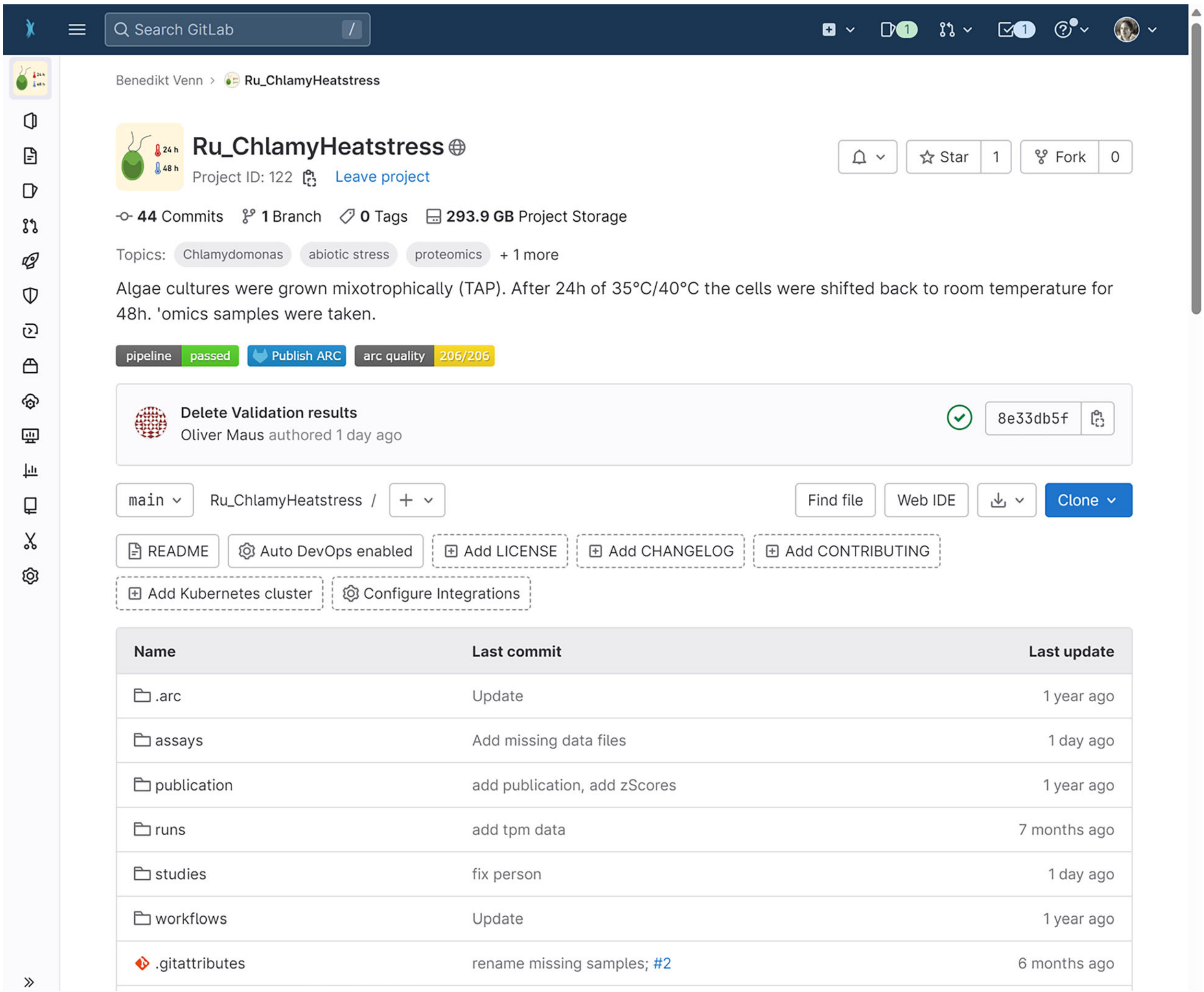


Figure 1. Schematic overview of the PLANTdataHUB platform. (A) A Keycloak instance acts as a gateway authorization and authentication infrastructure (AAI) unifying user identity across the hub instances. (B) The Git-based annotated research context (ARC) is stored and hosted a central GitLab instance that offers collaboration and project management tools. Additionally, there are many tools for working with ARCs. (C) ARC registry – a service for searching ARCs based on their metadata – acts across different hub instances. (D) ARCs can be exported to other endpoint repositories or to a connected InvenioRDM instance which issues DOIs for publication.



Ru_ChlamyHeatstress

Project ID: 122

44 Commits 1 Branch 0 Tags 293.9 GB Project Storage

Topics: Chlamydomonas abiotic stress proteomics + 1 more

Algae cultures were grown mixotrophically (TAP). After 24h of 35°C/40°C the cells were shifted back to room temperature for 48h. 'omics samples were taken.

Delete Validation results
Oliver Maus authored 1 day ago 8e33db5f

Ru_ChlamyHeatstress /

Name	Last commit	Last update
.arc	Update	1 year ago
assays	Add missing data files	1 day ago
publication	add publication, add zScores	1 year ago
runs	add tpm data	7 months ago
studies	fix person	1 day ago
workflows	Update	1 year ago
.gitattributes	rename missing samples; #2	6 months ago

Figure 2. View of a selected ARC on the PLANTdataHUB website user interface. The project view is a page that shows the details and features of a specific project. Depending on the project settings and permissions, you can see different information in the project view, such as: The project name, description, accessibility, branches, tags, commits, and merge requests. Below, the ARC directories and files are shown. On the menu bar on top there is project issues, boards, milestones, labels, and service desk.

or public, accessible to anyone. During the publication process, metadata is extracted from the ARC and seamlessly transferred to the InvenioRDM platform. This process enables the creation of a comprehensive DOI for the investigation in the form of a Git project. Additionally, a separate DOI is assigned to the current snapshot, facilitating ongoing research, data additions, and continuous project evolution until the next DOI is warranted. It is important to note that unlike in a standard commit scenario, the publication process cannot be triggered at any arbitrary state of an ARC. This restriction is in place to ensure data quality. To maintain this quality, an automated validation process for data publication in InvenioRDM is continuously running. This process generates a badge that reflects the quality of an ARC, signifying the availability of essential metadata and the validity of the data files associated with the ARCs (Figure 2). While most aspects are handled automatically, a final step of manual approval is necessary, which can be utilized to incorporate a community-driven review process if desired. This manual check serves as a final sanity check, conducted by humans. The automated quality control process, aimed at achieving compatibility and facilitating export to other endpoint repositories, serves as the technical blueprint.

While the central PLANTdataHUB (https://git.nfdi4plants.org) is ready for community use, the architecture supports a federated data hub landscape. This means that a containerized version of the customized GitLab can be downloaded (https://github.com/nfdi4plants/DataHUB) and run locally, supporting physical proximity interaction, and leveraging of existing hardware. This version is preconfigured to interact with the central Keycloak identity provider, enabling collaboration with individuals from other instances and access to the ARC registry, which allows searching across different instances and locations.

PLANTdataHUB is positioned at the core of plant research RDM, complemented by various tools integrated around the HUB to facilitate continuous annotation and FAIR collaboration using ARCs. These tools include free options from the extensive Git community and the DataPLANT NFDI4plants community, available for interaction either through the command line or a user-friendly interface. The data plant tool chain is supported by an extensive knowledge base (https://nfdi4plants.org/nfdi4plants.knowledgebase/index.html), including tutorials and documentation that can be readily extended by the community and an online tool to create ARCsupported data management plans (Zhou et al., 2023). The ARCitect (DataPLANT Community, 2023c), for instance, not only streamlines interactions with PLANTdataHUB but also simplifies ARC creation and management. Additionally, tools are available for ontology-driven metadata annotation of experimental and technical workflows. One such tool is a web-based add-in designed for both online and desktop versions of MS Excel. It reduces the reliance on free-text descriptions in metadata tracking, while still allowing comments and providing structural flexibility to extend metadata sheets. Compliance with ISA standards is achieved through headers and subsequent values derived from an ontology. Users can select applicable terms from a database of multiple ontologies, which can be browsed using a search function and expanded by the community. Furthermore, we offer over 40 metadata templates compatible with various endpoint repositories for direct export and various experimental workflows. These templates include technology-specific designs that align with the requirements of Minimal Standard Initiatives, such as MIAPPE (Papoutsoglou et al., 2020) for plant phenotyping as well as typical omics standards for submission to domain-specific providers such as NCBI, EBI, etc. Moreover, we actively engage the community in the development of lab or facility-specific workflows. Presently, 30% of the templates are contributed by the community, fostering collaboration, and ensuring that PLANTdataHUB remains a dynamic and evolving resource.

DISCUSSION

The FAIR data principles are driving data sharing, publication, and reuse in the research community. FAIR compliance is increasingly becoming a central requirement for various research infrastructures and communities. Additionally, research funding mandates are progressively demanding that projects make their data, models, and analyses adherent to FAIR principles. By facilitating FDO (FAIR Data Object) creation and data publication, PLANTdataHUB actively supports collaborative and continuous FAIR data management in the field of plant research.

On shifting plant research to a commit-based environment by using ARCs

While the usage of Git for managing research assets is not a new idea (Engwall & Roe, 2020; Ram, 2013; Vuorre & Curley, 2018), present approaches usually end up promoting the usage of Git via it's command line interface (CLI) directly for tracking changes to an unstructured set of assets. This does not only leave it up to the individual how to structure the content, but also requires the technical skill in using Git directly. ARCs in contrast are essentially Git repositories containing a well-defined, but flexible structure that is hidden behind user-friendly and mission-targeted interfaces.

Comparison with other repositories

The landscape of research data repositories is large and dynamic, as the focus of scientific communities increasingly shifts to making new and existing research data FAIR. Therefore, it is not feasible to attempt an exhaustive list of such repositories and we highlight a few 'large', state-of-the-art representatives of the space of repositories and compare different approaches and their implementations with our approach.

The general focus of already established multi-domain platforms such as Figshare (Singh, 2011) or Zenodo by OpenAIRE (European Organization For Nuclear Research & OpenAIRE, 2013) is to offer a frictionless archival of unstructured data of any kind, with the focus on making previously 'unpublishable' research assets such as figures, negative results, or datasets citable. Both platforms are widely used for that purpose, owing a large share of their popularity to their low-overhead approach. Strong emphasis lies on the first two FAIR data principles (Findability and Accessibility) by indexing records and assigning unique DOIs to them. However, not requiring any extensive (meta)data standards for the published assets is detrimental to the other principles (Interoperability and Reusability), as there can be no unified access to or reuse of fundamentally unstructured data. In addition, the amount of data to be submitted is usually limited. Omics discipline—specific archives like the NCBI short read archive or Metabolights allow submitting large data sets but necessitate deconstruction of experiments into individual domain-specific data sets that often lead to losing the connection between different measurements performed on the same sample.

In contrast, PLANTdataHUB tries to find the right balance between mandated structure and permissiveness by using ARCs as the fundamental format. The focus here is facilitating the shift to a fundamentally versioned workflow for scientific projects from start to finish across all experimental data. Here, FAIR Data Objects naturally emerge over the course of a project, therefore offloading the mandated overhead of using the ARC structure. This is further supported by productivity gains via project management tools and Continuous Quality Control. While annotating provenance of an already existing single figure, negative result, or dataset are all possible within an ARC after the fact, the PLANTdataHUB platform is better suited when it accompanies the actual genesis of such assets. Therefore, Figshare or Zenodo remain solid choices for making existing assets citable without explicitly tracking their provenance (e.g., because they are already part of a publication).

FAIRDOM-SEEK (Wolstencroft et al., 2017) is a platform that also uses an extended ISA format for (meta)data annotation with ontologies, an approach very similar to how ARCs employ the standard. Flow (Capitanchik et al., 2023) on the other hand is an upcoming platform that focuses on workflow execution. It employs its own metadata scheme that is continuously verified in a similar manner to PLANTdataHUB's Continuous Quality Control. PLANTdataHUB is set apart from both platforms by providing a collaborative environment that supports the scientific research process. In direct comparison with FAIRDOM SEEKs metadata standard, the difference lies in the mandated folder structure of the ARC and its underlying Git-based concepts, the benefits of which were exhaustively discussed. While FAIRDOM SEEK has integrations for running workflows, they are offloaded to an external platform exclusively instead of additionally offering a direct runner platform like PLANTdataHUB or Flow. Using the existing, proven exhaustive and adaptable ISA standard, on the other hand, has obvious advantages in contrast to designing a new metadata format, which comes with a plethora of challenges.

Finally, there are FAIR Data Object aggregation engines, such as re3data (Pampel et al., 2013) or OpenAIRE Explore (Rettberg & Schmidt, 2012). PLANTdataHUB is a platform for ARCs exclusively and can therefore serve as a source of well-curated research data for these engines. While PLANTdataHUB offers a federated cross-instance index for ARCs, it does not aim to aggregate other metadata standards directly. However, it is planned to support automated conversion from other standards into ARCs in the future.

In summary, PLANTdataHUB provides a collaborative data curation and publication approach, a community-driven knowledgebase, and data resource for plant scientists.

EXPERIMENTAL PROCEDURES

Annotated research context overview

The PLANTdataHUB platform gathers ARCs as an independent concept to store and organize research data. The ARC specification is based on lightweight principles to organize elements of data-driven research (data, metadata, computational workflows, and results) in files and directories in a specific, versioned layout (Figure 3). It leverages multiple open file formats and other standards to ensure interoperability and reproducibility of the information and elements contained within each ARC. The fundamental design goal of ARCs is twofold. On one hand, ARC principles assist plant researchers in organizing their data such that they are enabled to maximally profit from existing standards, tools, and infrastructures. On the other hand, ARCs are FDOs, thus working with ARCs ensures FAIRness of research.

Data and metadata standards

Data and metadata annotations are performed according to the widely used ISA (Investigation, Study, Assay) model (Sansone et al., 2012). While originally designed for omics metadata, ISA descriptions have been repurposed and pervasively applied for other data types that may be contained in ARCs. This allows the use of a wide variety of tools from the ISA ecosystem with ARCs to support researchers in typical metadata annotation tasks. The ISA specification defines an abstract metadata model and requires annotation of specific ontology terms at the Investigation, Study, and Assay levels: the *Investigation file* is unique per ARC and contains the top-level metadata, such as personal information about the contributing researchers or a description of the project. *Study files* track the provenance of resources used in subsequent assays, for example, by annotating the sample extraction protocols used to create a given sample or the repository where an input file was retrieved from. *Assay files* then annotate the genesis of measurement data, for example by annotating the parameters of a next-generation sequencing run or the arguments used for a data analysis tool. Additional terms from other ontologies (e.g., to annotate units, protocols, and data formats) are easily included if considered useful for specific data types or research questions. ISA metadata is stored in a tabular format (ISA-XLSX) that is easy to edit using typical spreadsheet applications (Figure 4).

Similarly, computational workflows within ARCs are represented using the *Common Workflow Language* (CWL, cf. Figure 5) (Crusoe et al., 2022), enabling ARCs to interoperate within a large ecosystem of computational environments [e.g., in the de.NBI cloud (Belmann et al., 2019; Pühler, 2016), on Galaxy instances (The Galaxy Community, 2022), etc.] and computational tools [e.g., BioCompute objects (Simonyan et al., 2017), etc.]. While other workflow languages are also in widespread use [e.g., Nextflow (Di Tommaso et al., 2017) or snakemake (Köster & Rahmann, 2012)], CWL wrapping can be used to achieve a compatibility layer. Through CWL's container facilities, ARC users can easily access a plethora of containerized computational tools for analysis and visualization.

Versioning

PLANTdataHUB is built upon the core idea that scientific progress thrives on dynamic and ever-evolving information, while measurement data remains unalterable. This fundamental premise leads to a perspective where minor, step-by-step modifications can continually stack upon one another, shaping an ever-evolving yet immutable research entity.

Taking inspiration from principles and processes typically used in software development, PLANTdataHUB suggests that scientific advancement can be fundamentally linked to a compilation of discrete, self-contained incremental adjustments – *commits* – to a project, for example, to an ARC. Commits are comprised of the changes made and a description that can add additional context

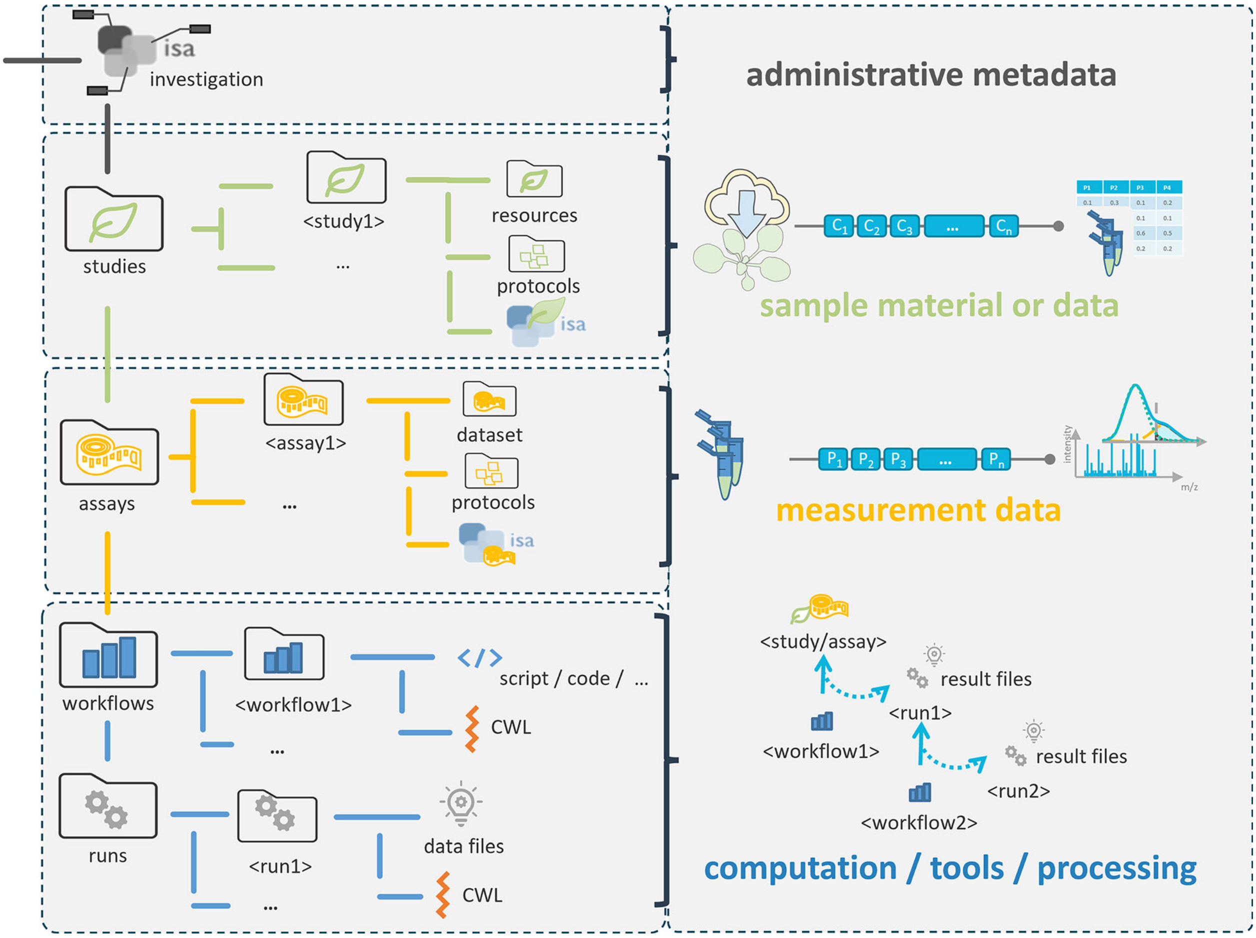
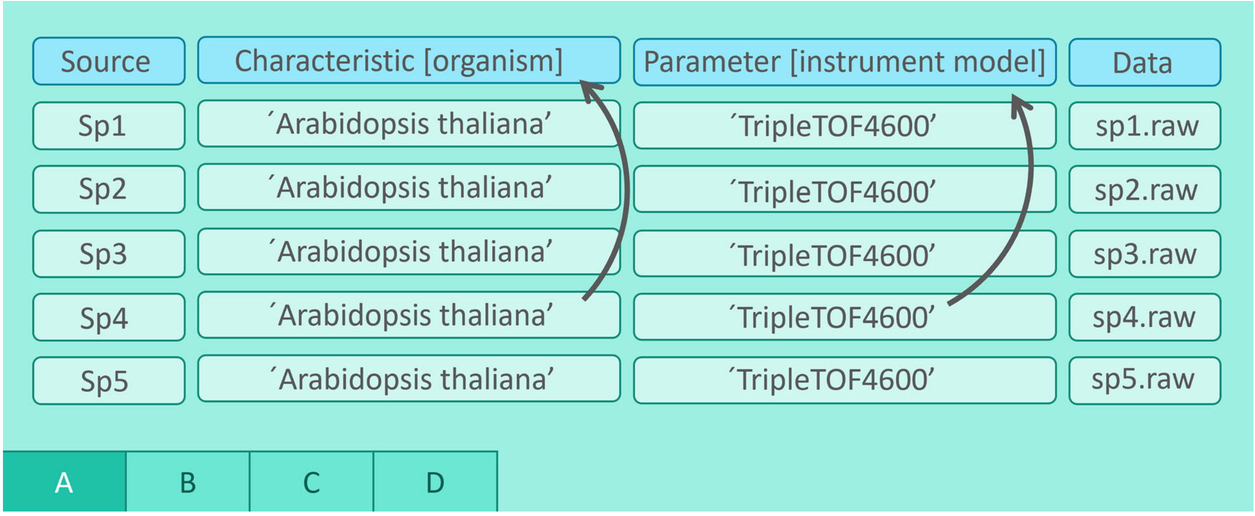


Figure 3. Folder and file layout in annotated research contexts (ARCs). The top-level investigation metadata file (*isa.investigation.xlsx*) contains administrative metadata about the ARC, for example, the names and affiliations. The *studies* folder (green) contains one or multiple study subfolders, which are used to track provenance of the contained resource and protocol files, as well as study metadata files that annotate the genesis of biological samples and other material. Analogously, the *assays* folder (yellow) contains one or multiple assay subfolders, which are used to track provenance of the contained datasets and protocol files, as well as assay metadata files that annotate the genesis of measurement data and other result files starting from study material or datasets. The *workflows* folder (blue) contains common workflow language (CWL) descriptions of, for example, analysis scripts or tools, while the *runs* folder contains CWL descriptions of the actual analysis that was performed, making the computational pipeline used in the ARC reproducible.

to those changes (Figure 6). Over the course of a scientific project, insights usually accumulate and cross-reference over time, which is naturally captured in a commit-based representation. Even when the history of a project is not of primary interest while it is being conducted, commits effortlessly hold value as they document the process of how the project came to its conclusion. Additionally, having a history of commits can be immensely useful in drafting manuscripts to present results. Fine-grained commits allow re-analysis and reevaluation of data, for example, to examine transcriptomics experiments with differing genomic annotation.

Mirroring this reality, ARCs are explicitly versioned and are represented as Git (Chacon, 2014) repositories. Git is an established open-source version control system (VCS) that tracks changes to files in a directory, and is therefore directly applicable to the ARC, which is represented by such a structure. As motivated above, changes between ARC versions are represented by Git commits, which record changes to an arbitrary number of files in an ARC, automatically including all of an ARCs data and metadata in the versioning. In addition to changes in files, additional provenance data, such as author information and commit messages, can be captured in the ARC repository's version history, affording a large measure of transparency. Afforded by the combination of lightweight, user-friendly annotation facilities and fine-grained versioning capabilities, ARCs can be treated as evolving collections of research data that are continuously extended, annotated, curated, or otherwise improved (Garth et al., 2021). This contrasts with other format and layout guidelines or specifications [e.g., RO Crate (Sefton et al., 2023)], which are first and foremost designed to accommodate archival needs, thus advocating an immutable view of research data that is not conducive to everyday RDM needs.

Starting from an empty ARC, the sequence of commits in an ARC forms a coherent history that tracks the provenance of all its contents. A commit history is not necessarily linear, supporting branching out of individual trains of thought and subprojects that can be either discarded or incorporated into the main history, therefore addressing the dynamic nature of research. This manner



Source	Characteristic [organism]	Parameter [instrument model]	Data
Sp1	´Arabidopsis thaliana’	´TripleTOF4600’	sp1.raw
Sp2	´Arabidopsis thaliana’	´TripleTOF4600’	sp2.raw
Sp3	´Arabidopsis thaliana’	´TripleTOF4600’	sp3.raw
Sp4	´Arabidopsis thaliana’	´TripleTOF4600’	sp4.raw
Sp5	´Arabidopsis thaliana’	´TripleTOF4600’	sp5.raw

Figure 4. Example of an ISA annotation table, showing expressive annotation guided by the hierarchy of ontologies (represented by arrows). Each row is interpreted left to right and describes the metadata of the (exemplary) data.

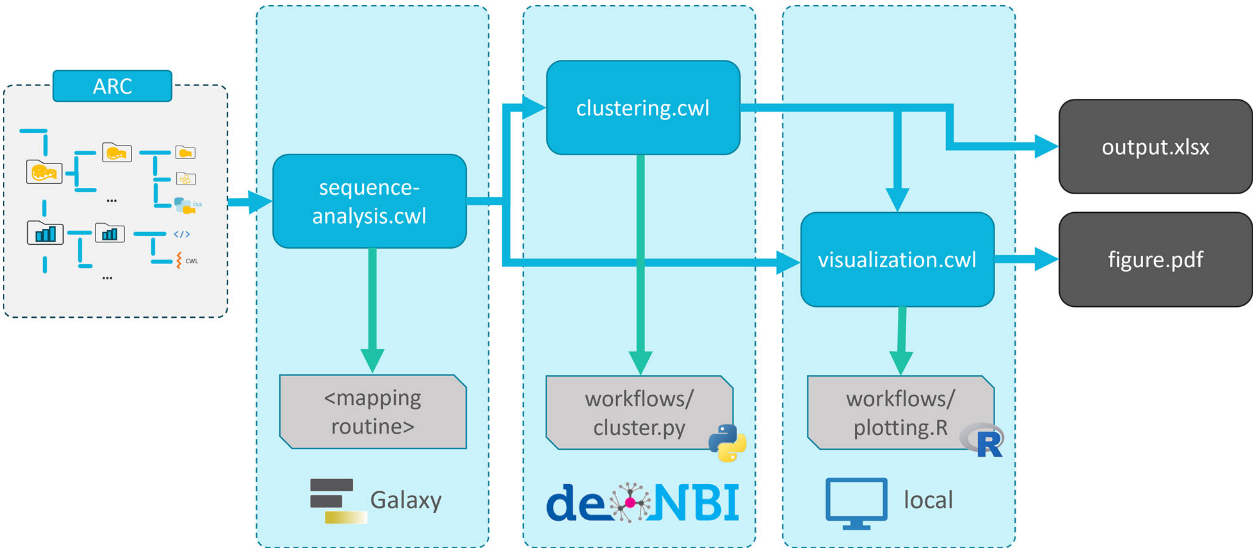


Figure 5. Workflows describe and automate computational analysis. While CWL workflow descriptions reside in ARCs, computation steps can target a wider set of computational resources resulting in joint result data files.

of versioning – based on a fundamental unit of change – respects the complex ever-evolving nature of an ARC while also transparently documenting not only data provenance but also the genesis of the contained scientific arguments. The project state can always be monitored and evaluated based on the commits added since the previous evaluation session. A fundamentally versioned work environment also means that introductions of errors can be tracked based on its history. Planning a specific sampling procedure means committing the selected protocol with the chosen parameters to the experimental metadata of the project. Changing a single parameter within that protocol, adding the results of an analysis script, or improving the description of the aims of a project are all commits, as are re-organizing the folder structure that contains results or adding the name of a collaborator to the ARCs metadata.

Technically, ARC versions can be managed at a low level using any Git implementation confirming to the open standard, for example, the git CLI tool or a GUI such as SourceTree (Atlassian, 2023); however, a user-friendly level of interaction specifically developed for plant researchers can be delivered by dedicated tools such as the ArcCommander (DataPLANT Community, 2023b) and ARCitect (DataPLANT Community, 2023c) that is available as part of the PLANTdataHUB tool chain. Since Git is fundamentally based on hashing, file, and data integrity are automatically ensured and provenance can be proven if needed. Beyond simply representing the state of research at a given point in time, every Git repository, and hence every ARC, contains its own full history. Large data files such as many raw data files in the omics disciplines are transparently accommodated via the Git Large File Storage extension (Atlassian

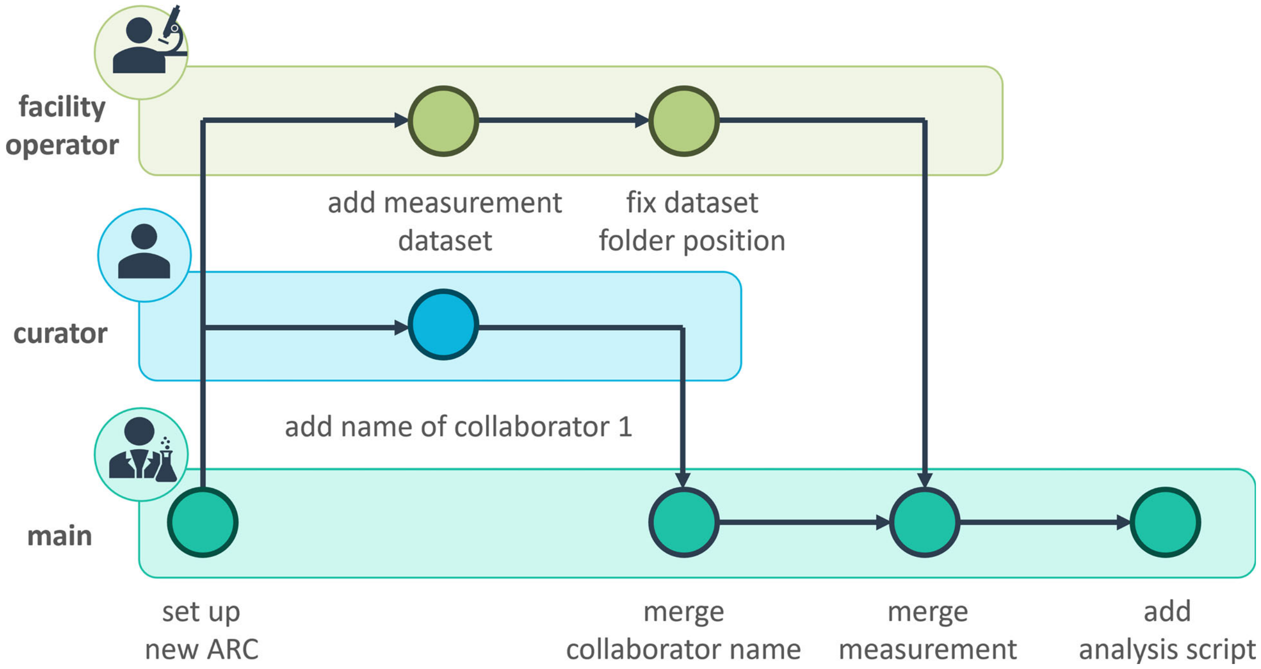


Figure 6. Schematic ARC commit history. In a straightforward commit-based work environment, there are three branches that eventually merge into the main branch. Commits are represented as circles. Initially, after the first commit labeled 'Set up new ARC' in the default main branch (bottom branch), two new branches are spawned, namely "curator" and "facility operator." Within the "curator" branch, a solitary commit is made to include a collaborator's name in the ARC metadata. In contrast, within the "facility operator" branch, two commits are introduced to insert a dataset file into the ARC and rectify its placement. Upon incorporating the commits from both branches, work resumes directly on the main branch, featuring a commit that introduces an analysis script.

et al., 2023), which is a *de facto* standard and included in most Git-aware software tools.

Moving beyond the needs of a single researcher, version control is the basis for collaborative development of software. Git mechanisms such as forking and merging in combination with decentralized repositories enable granular collaboration mechanisms. Fundamentally leveraging their Git heritage, ARCs are easily used in collaborative research. Furthermore, collaboration requires communication; based on versioned ARCs, PLANTdataHUB offers a wide array of project management and automation tools that reap the benefits of the versioning approach, providing a platform that efficiently stores, hosts, and facilitates collaborative work with them.

In contrast to proprietary platforms such as GitHub (Perkel, 2016), PLANTdataHUB does not enforce collaboration to occur forcibly through a central resource. For example, a dedicated institute instance can facilitate collaboration on premises, without forcing researchers to share their data to an external provider.

The PLANTdataHUB

The core component of the PLANTdataHUB platform is a GitLab (Engwall & Roe, 2020; GitLab, 2023) instance that stores ARCs. GitLab is a Development Operations platform. The latter ensures the integration of software development and information technology operation processes with strong automation (Ebert et al., 2016). By using and expanding the tools provided by GitLab, PLANTdataHUB transcends typical data archives and assists researchers during the entire cycle of scientific research from project management to data management and quality control, from collaborative writing to publication of ARCs.

The PLANTdataHUB platform is not only a specific infrastructure, but also a blueprint from which further instances can be created via provisioning and container technologies. The blueprint scales well to available resources and can be employed both on small scales (such as a PC running in a research group) and large-scale infrastructures (such as bioinformatics clouds). The DataPLANT consortium operates the nationally supported PLANTdataHUB instance that is available to all plant researchers as a resource.

Project management and collaboration

Scientific research projects are more and more conducted in collaboration among multiple researchers, often distributed across different facilities, institutions, or countries, focusing on different aspects of an investigation, and progressing individually. Without mechanisms and a platform facilitating low-friction collaboration, coordination of these efforts can be a complicated and frustrating task that often defaults to the usage of individual correspondence and file-sharing services, which are neither fault-tolerant nor scalable.

As stated above, PLANTdataHUB borrows project management tools and collaborative processes that are typically used in software development, such as version controlling and maintaining different versions (Figure 7). For example, when needed, the commit history of ARCs can split into *branches* - that represent independent developments starting from the next commit after branching out. ARCs usually have a canonical main branch, from which independent branches can be formed and be merged if needed.

Consider the following scenario: at a certain point in the main branch, the principal investigator might decide to involve a collaborator to analyze a dataset with their signature method. The collaborator can then branch out from the main branch and work on their analysis without impairing the further work done on the

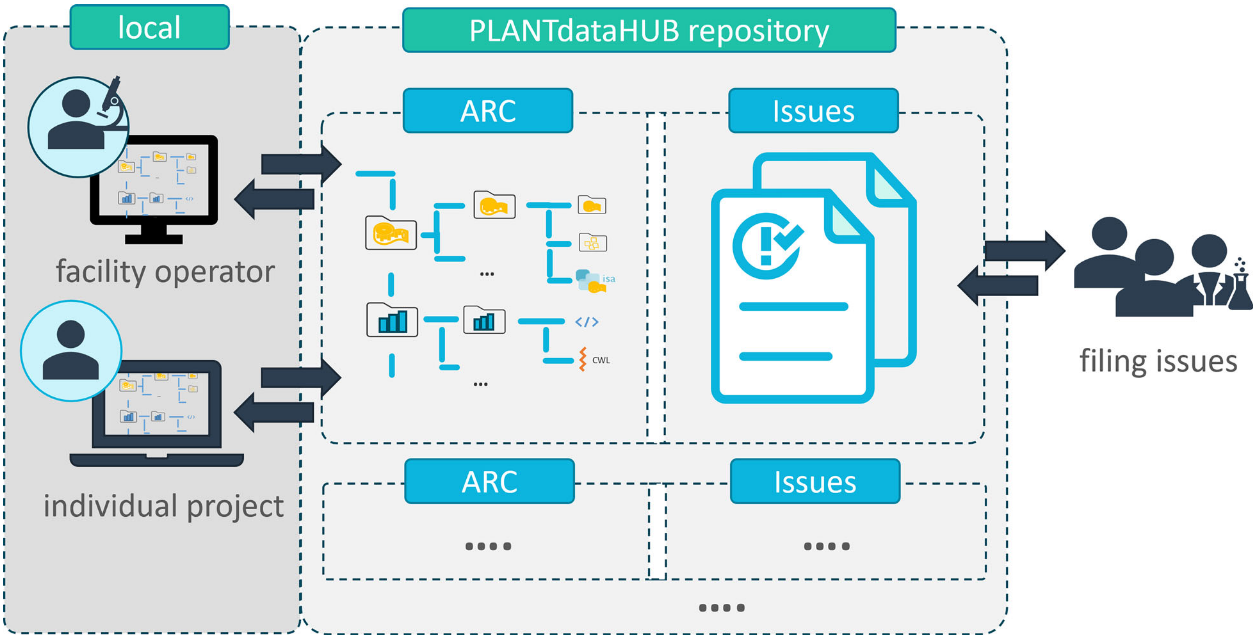


Figure 7. Collaborative research projects. Each PLANTdataHUB instance provides a central ARC repository per research project, together with project management facilities (issue management). Collaboration only requires that all participants contributing to a project agree on a particular hub instance, for example, the central instance, or a custom instance hosted at an institute or in a workgroup.

main branch of the project. Once their analysis is done, they can get their independent work merged into the main branch. Once the analysis branch is merged, it is also documented who performed the analysis, as commits are linked to the user that created them. The ARC directory structure is designed to anticipate distributed contributions. For example, a facility can a particular measurement by placing the dataset into a separate assay, extending an existing ARC that already contains study design and sample extraction.

In analogy to the atomic commit as the basis for RDM within ARCs, the core of PLANTdataHUB project management is again an atomic unit – the *issue* – which describes a unit of work necessary to accomplish a particular improvement. Issues can be: a reminder to improve a paragraph in a manuscript, a discussion on how to analyze a dataset, or a review comments. Similarly, questions from other researchers regarding methodology of a published ARC can be represented in issues. Issues always belong to a single project (i.e., ARC) and have a discussion thread associated with them that is accessible for everyone with permission to view the ARC.

Several ARCs can be organized as a group, for which all issues can be viewed at once. This is especially useful for researchers managing multiple projects or facility managers. Custom tags can be used to further categorize issues, which can then be incorporated into advanced project management tools such as Kanban boards (Corona & Pani, 2012)– a project state visualization technique that depicts various stages of the project as columns containing cards representing individual work items (issues). Combined with different access control (AC) systems, these tools enable virtually any collaborative model, from a single person planning a project on a private ARC that is exclusively visible to them to a published multi-ARC group that is available for everyone.

Access control is paramount for not penalizing researchers that strive for FAIR data with mandatory transparency in early project stages, which would lead to only finished projects being uploaded to PLANTdataHUB instances and hamper free discussion in issues. Therefore, several access control (AC) systems for the ARCs on the instance are available: AC via *visibility* simply sets read-only access to a resource either for everyone (*public*, visible for non-authenticated users), all authenticated users on the instance (*internal*), or selected authenticated accounts (*private*). *Project* and *Group*-based AC on the other hand work by assigning *roles* for a set of people for either a single project or group of projects. These roles control the type of access to the target resources, for example a *reporter* basically only has access to related project management tools, while a *maintainer* can make all non-destructive changes to the resource (e.g., create a commit, but not delete the ARC).

Quality control and automation opportunities

Quality control of research data – for example, (meta)data integrity checks or schema conformity for target archives – is usually most rigorously applied when it seems immediately useful, for example, directly before submitting data to a repository or submitting a manuscript to a journal. Paradoxically, this can be the stage of a project where failing these checks causes the most damage, for example by missing a deadline for submitting a manuscript because mandatory data is not available or worse yet corrupted. Failing to perform proper quality control at all can even lead to rejection or retraction of the work. This checkpoint-based submission of data, which then is usually treated as 'done' after submission, is inflexible and therefore often seen as a nuisance and not more than a checkbox required by journals.

In addition to 'manual' quality control via raising issues, discussing possible solutions and how to implement them, or reviewing commits before they are added to the history via project management tools, the incremental changes arising in a commit-based model such as the ARC provides another striking benefit: *continuous quality control (CQC)*. Upon every commit to an ARC repository, PLANTdataHUB can perform automated quality control measures based on the committed changes.

For example, when a commit is performed adding a new dataset to an ARC, PLANTdataHUB can automatically check whether the metadata in the ARC describing the new dataset is reasonably complete. If the check fails, the hub can trigger a response that is deemed appropriate given the project context. For example, an issue can be automatically created to notify a contributor of missing (meta)data.

There are many use cases and scenarios for quality control of research data. Beyond allowing implementation of an abstract measure of quality, an immediate practical application in plant biology is to ensure that it is always possible to export project data and metadata into another format to deposition into a research repository, such as the Arabidopsis information resource TAIR (Rhee et al., 2003). Having this check performed automatically and continuously on each new commit means that researchers working on this ARC are assured that their data adheres to the selected quality parameters. If this is not the case, it is straightforward to identify when (i.e., at which commit) the ARC became non-compliant, allowing us to quickly address the problem(s).

Beyond CQC, the automation mechanisms included in PLANTdataHUB can be used for a plethora of other use cases. For example, it is possible to automate computational analysis based on the state of data and metadata in an ARC, leveraging ARC's built-in reproducibility mechanisms that describe computational analysis pipelines described in CWL files (Figure 5). After each commit, PLANTdataHUB can be set up to automatically re-run some or all analyses described in an ARC. In this manner, PLANTdataHUB can also act as an analysis platform, especially for common pipelines and variable input. Moreover, PLANTdataHUB can also be configured to automatically tell other platforms or infrastructures about the changes in an ARC. For example, this allows indexing of metadata for ongoing projects in searchable databases, where the database is always up-to-date with the project state.

CQC and further automation are implemented by once again relying on existing GitLab infrastructure, namely the automation pipeline runner architecture. A basic integrity check (e.g., whether a person in the ARC metadata has a name) is performed for all ARCs on the instance on each commit. Researchers can create additional automation steps by adding a configuration file to the ARC and either reference pre-existing pipelines or create their own. Automation jobs for a commit are then queued and eventually picked up and performed by a runner – an application that executes the job in the context of the ARC at the current commit. Runners can be co-located on the same hardware as the PLANTdataHUB but can target large public infrastructures such as the de.NBI cloud for use cases involving large data. Alternatively, analysis platforms can be extended to interact with ARCs in the PLANTdataHUB; for example, GALAXY is able to leverage ARCs as a source of data.

Data publication

Data is usually the primary output of scientific research; in contrast, the universal measurement of research success is the publication. In the past, accessing the underlying data of a published research paper required the tedious process of contacting the original authors and asking for access. With the continuing rise of high-throughput methodologies and the need for meta-analysis, datatype-specific repositories have seen an increase in popularity. For example, it is mandatory to deposit raw sequence reads of a next-generation sequencing experiment on one of the Sequence Read Archive mirrors (Cochrane et al., 2016) alongside the publication for many journals. If researchers include this data in a new publication, they must cite the publication, because although the dataset has a unique database identifier, no canonically citable identifier [such as a Digital Object Identifier (DOI) (Paskin, 2009)] that directly points to it is available. As an unintended consequence of this approach, data that is not included in a publication is not FAIRly citable. Hence, it cannot produce academic value in

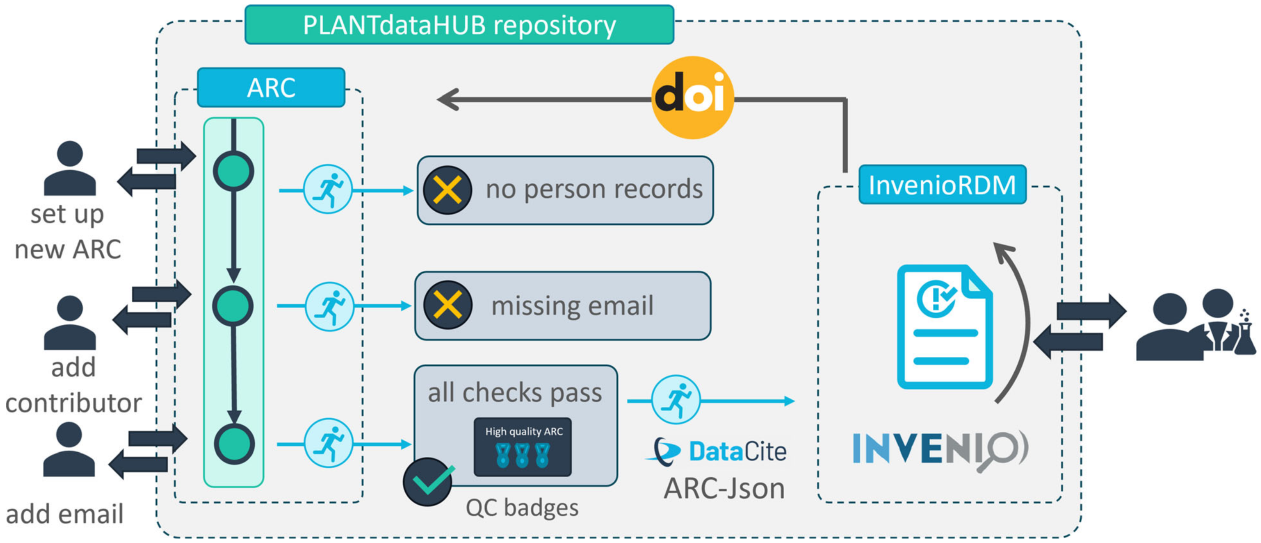


Figure 8. Publication flow - continuous quality control in a commit-based workflow. Circles depict commits as nodes in the commit history. Changes trigger a pipeline runner – a program that executes other quality control checking programs in the context of a given commit – to perform quality control measures and return a report. If all checks are passed, the ARC can be considered eligible for publication. The publication process inserts a record to InvenioRDM, which will then issue a Digital Object Identifier.

the form of citations for the original authors, although it might be of use in meta-analyses or follow-up investigations.

Individual PLANTdataHUB instances (such as the one operated by DataPLANT) can be connected to archiving software that can link eligible ARCs to permanent metadata records. Each ARC on such an instance can have quality information indicators (e.g., badges) that inform the user about the quality of its metadata and if the necessary metadata for creating a data publication is available. If that is the case, the publication process can be initiated for the ARC.

In technical implementation (cf. Figure 8), the central PLANTdataHUB is connected to an InvenioRDM representing the underlying technology for the well-known Zenodo data repository (Invenio Community, 2023; Kaplun, 2010) instance that acts as interface to DataCite (Brase, 2009), which issues the final DOIs. ARCs that pass a CQC pipeline with the following steps are eligible for a data publication: (i) a machine-readable representation (JSON) of the ARC metadata is created and linked with the commit, which makes it discoverable for search tools. (ii) ARC metadata is subjected to a set of quality checks, for example, if ORCIDs associated with persons are valid or if each author has a contact email. (iii) The ARC metadata is converted to a metadata record, which is submitted to the InvenioRDM instance that in turn triggers the creation of a DOI via DataCite for that record.

Federation and integrative tools

To allow PLANTdataHUB instances to move beyond a role as isolated information silos, the platform is designed to enable a federation, that is, data in an instance hosted by the users' university is findable in the central PLANTdataHUB, if the user specifically opts into sharing their data. Beyond providing increased opportunities for data sharing, collaboration, and dissemination, federation with other instances carries further advantages. Users are able to deploy custom PLANTdataHUB instances (e.g., institutions) instead of using the DataPLANT-provided central instance, while still retaining the ability to opt-in to these further services.

User authentication and authorization

Managing credential data for the ever-increasing amount of online services can be a tedious burden on researchers and lab or institution administrators. Therefore, DataPLANT offers a federated authentication system that all PLANTdataHUB instances can choose to integrate with it. Users can use their existing and widely-used identity providers such as ORCID (Haak et al., 2012) or Life Science login (European Life Science Research Infrastructures, 2023), or choose to create a new account that is associated with an email address. DataPLANT's login federation service then handles connecting users' accounts on all hub instances. Therefore, responsibility for user management can be fully delegated to DataPLANT.

The heart of DataPLANT's login federation system is Keycloak (Christie et al., 2017; Keycloak - Identity and Access Management for Modern Applications, 2021), an open-source identity and access management system that acts as a single-sign-on proxy for external authentication providers. This setup provides support for virtually any login method, as it maps from the authentication provider to an internal Keycloak account, which is then used to authenticate to the federated PLANTdataHUB instances. This can be extended to include other authentication/authorization providers (e.g., social login) should the need arise.

Indexing

Federated PLANTdataHUB instances can also choose to take part in DataPLANT's ARC registry, which improves findability and accessibility of ARCs stored in the instance. The ARC registry is a cross-instance service that provides an advanced search interface for ARCs based on their ISA-formatted metadata. Using this service, it is possible to query multiple layers of metadata across all indexed ARCs: Investigation metadata, for example, all ARCs where a certain person is on the author list, Study metadata, for example, all ARCs that used material sampled from a certain organism, and Assay metadata, for example, all ARCs that performed next generation sequencing with the instrument model of interest. Queries that search for combinations of these levels are possible as well. The ARC registry is especially useful for meta-analysis, as it leverages the ontology-based metadata annotations.

ACKNOWLEDGEMENTS

All individual contributors are credited on the DataPLANT website. We thank the plant community for their ever-growing engagement towards FAIR. Open Access funding enabled and organized by Projekt DEAL.

FUNDING INFORMATION

This work was supported by the German Research Foundation (DFG) within the Project "Establishment of the National Research Data Infrastructure (NFDI)" in the consortium DataPLANT (442077441) and the TRR 175 (INF) (270050988).

REFERENCES

Atlassian. (2023) Sourcetree|Free Git GUI for Mac and Windows. SourceTree. Available from: https://www.sourcetreeapp.com [Accessed 31st August 2023]

Atlassian, GitHub & Open Source Contributors. (2023) Git Large File Storage. Available from: https://git-lfs.com/ [Accessed 31st August 2023]

Barrett, T. & Edgar, R. (2006) Gene expression omnibus: microarray data storage, submission, retrieval, and analysis. In: Alan, K. & Brian, O. (Eds.) *Methods in enzymology, DNA microarrays, part B: databases and statistics*. Cambridge, MA, USA: Academic Press, pp. 352–369. Available from: https://doi.org/10.1016/S0076-6879(06)11019-8

Belmann, P., Fischer, B., Krüger, J., Procházka, M., Rasche, H., Prinz, M. *et al.* (2019) de.NBI Cloud federation through ELIXIR AAI. *F1000Research*, **8**, 842. Available from: https://doi.org/10.12688/f1000research.19013.1

Brase, J. (2009) DataCite - A Global Registration Agency for Research Data. In: Randall, B. (Ed.) *2009 Fourth International Conference on Cooperation and Promotion of Information Resources in Science and Technology. Presented at the 2009 Fourth International Conference on Cooperation and Promotion of Information Resources in Science and Technology*. Piscataway, NJ, USA: IEEE, pp. 257–261. Available from: https://doi.org/10.1109/COINFO.2009.66

Brickley, D., Burgess, M. & Noy, N. (2019) Google Dataset Search: Building a search engine for datasets in an open Web ecosystem. In: Ling, L. & Ryen, W. (Eds.) *The World Wide Web Conference, WWW '19*. New York, NY: Association for Computing Machinery, pp. 1365–1375. Available from: https://doi.org/10.1145/3308558.3313685

Capitanchik, C., Ireland, S., Harston, A., Cheshire, C., Jones, D.M., Lee, F.C.Y. *et al.* (2023) Flow: a web platform and open database to analyse, store, curate and share bioinformatics data at scale. https://doi.org/10.1101/2023.08.22.544179

Chacon, S. (2014) *Pro Git, Second edition. The expert's voice in software development*. New York, NY: Apress.

Christie, M.A., Bhandar, A., Nakandala, S., Marru, S., Abeysinghe, E., Pamidighantam, S. *et al.* (2017) Using Keycloak for Gateway Authentication and Authorization. https://doi.org/10.6084/m9.figshare.5483557.v1

Cochrane, G., **Karsch-Mizrachi, I.**, **Takagi, T.** & **International Nucleotide Sequence Database Collaboration**. (2016) The International Nucleotide Sequence Database Collaboration. *Nucleic Acids Research*, **44**, D48–D50. Available from: https://doi.org/10.1093/nar/gkv1323

Corona, E. & **Pani, F.E.** (2012) An investigation of approaches to set up a Kanban board, and of tools to manage it. In: *Proceedings of the 11th International Conference on Telecommunications and Informatics, Proceedings of the 11th International Conference on Signal Processing, SITE'12*. Stevens Point, WI: World Scientific and Engineering Academy and Society (WSEAS), pp. 53–58.

Crusoe, M.R., **Abeln, S.**, **Iosup, A.**, **Amstutz, P.**, **Chilton, J.**, **Tijanić, N.** ***et al.*** (2022) Methods included: standardizing computational reuse and portability with the Common Workflow Language. *Communications of the ACM*, **65**, 54–63. Available from: https://doi.org/10.1145/3486897

DataPLANT Community. (2023a) Annotated Research Context Specification, v1.1-rfc. https://doi.org/10.5281/zenodo.8302662

DataPLANT Community. (2023b) ARC Commander.

DataPLANT Community. (2023c) nfdi4plants/ARCitect: Arcitect. https://doi.org/10.5281/zenodo.8307729

Di Tommaso, P., **Chatzou, M.**, **Floden, E.W.**, **Barja, P.P.**, **Palumbo, E.** & **Notredame, C.** (2017) Nextflow enables reproducible computational workflows. *Nature Biotechnology*, **35**, 316–319. Available from: https://doi.org/10.1038/nbt.3820

Ebert, C., **Gallardo, G.**, **Hernantes, J.** & **Serrano, N.** (2016) DevOps. *IEEE Software*, **33**, 94–100. Available from: https://doi.org/10.1109/MS.2016.68

Engwall, K. & **Roe, M.** (2020) Git and GitLab in library website change management workflows. *The Code4Lib Journal*.

European Life Science Research Infrastructures. (2023) LS Login¦LifeScience RI. Available from: https://lifescience-ri.eu/ls-login/ [Accessed 31st August 2023]

European Organization For Nuclear Research & OpenAIRE. (2013) Zenodo: Research. Shared. https://doi.org/10.25495/7GXK-RD71

Garth, C., **Lukasczyk, J.**, **Mühlhaus, T.**, **Venn, B.**, **Krüger, J.**, **Glogowski, K.** ***et al.*** (2021) Immutable yet evolving: ARCs for permanent sharing in the research data-time continuum. In: Heuveline, V. & Bisheh, N. (Eds.) *HeiBOOKS E-Science-Tage 2021: Share Your Research Data*. Heidelberg: heiBOOKS, pp. 366–373. Available from: https://doi.org/10.11588/heibooks.979.c13751

GitLab. (2023) GitLab: The DevSecOps Platform. Available from: https://about.gitlab.com/ [Accessed 31st August 2023]

Haak, L.L., **Fenner, M.**, **Paglione, L.**, **Pentz, E.** & **Ratner, H.** (2012) ORCID: a system to uniquely identify researchers. *Learned Publishing*, **25**, 259–264. Available from: https://doi.org/10.1087/20120404

Haug, K., **Cochrane, K.**, **Nainala, V.C.**, **Williams, M.**, **Chang, J.**, **Jayaseelan, K.V.** ***et al.*** (2020) MetaboLights: a resource evolving in response to the needs of its scientific community. *Nucleic Acids Research*, **48**, D440–D444. Available from: https://doi.org/10.1093/nar/gkz1019

Haug, K., **Salek, R.M.** & **Steinbeck, C.** (2017) Global open data management in metabolomics. *Current Opinion in Chemical Biology*, **36**, 58–63. Available from: https://doi.org/10.1016/j.cbpa.2016.12.024

Hermjakob, H. & **Apweiler, R.** (2006) The Proteomics Identifications Database (PRIDE) and the ProteomExchange Consortium: making proteomics data accessible. *Expert Review of Proteomics*, **3**, 1–3. Available from: https://doi.org/10.1586/14789450.3.1.1

Invenio Community. (2023) InvenioRDM. — inveniosoftware.org. Available from: https://inveniosoftware.org/products/rdm/ [Accessed 31st August 2023]

Kaplun, S. (2010) Invenio: A Modern Digital Library System. The 5th International Conference on Open Repositories (OR2010), Madrid, Spain. https://doi.org/10.2390/biecoll-OR2010-10

Keycloak - Identity and Access Management for Modern Applications. (2021) *Keycloak - Identity and Access Management for Modern Applications*, 1st edition. Birmingham, UK: Packt Publishing.

Köster, J. & **Rahmann, S.** (2012) Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics*, **28**, 2520–2522. Available from: https://doi.org/10.1093/bioinformatics/bts480

Krantz, M., **Zimmer, D.**, **Adler, S.O.**, **Kitashova, A.**, **Klipp, E.**, **Mühlhaus, T.** ***et al.*** (2021) Data management and modeling in plant biology. *Frontiers in Plant Science*, **12**, 717958.

Krishnakumar, V., **Hanlon, M.R.**, **Contrino, S.**, **Ferlanti, E.S.**, **Karamycheva, S.**, **Kim, M.** ***et al.*** (2015) Araport: the Arabidopsis information portal. *Nucleic Acids Research*, **43**, D1003–D1009. Available from: https://doi.org/10.1093/nar/gku1200

Lawrence, C.J., **Dong, Q.**, **Polacco, M.L.**, **Seigfried, T.E.** & **Brendel, V.** (2004) MaizeGDB, the community database for maize genetics and genomics. *Nucleic Acids Research*, **32**, D393–D397. Available from: https://doi.org/10.1093/nar/gkh011

Mayer, G., **Müller, W.**, **Schork, K.**, **Uszkoreit, J.**, **Weidemann, A.**, **Wittig, U.** ***et al.*** (2021) Implementing FAIR data management within the German Network for Bioinformatics Infrastructure (de.NBI) exemplified by selected use cases. *Briefings in Bioinformatics*, **22**, bbab010. Available from: https://doi.org/10.1093/bib/bbab010

Pampel, H., **Vierkant, P.**, **Scholze, F.**, **Bertelmann, R.**, **Kindling, M.**, **Klump, J.** ***et al.*** (2013) Making research data repositories visible: the re3data.org Registry. *PLoS One*, **8**, e78080. Available from: https://doi.org/10.1371/journal.pone.0078080

Papoutsoglou, E.A., **Faria, D.**, **Arend, D.**, **Arnaud, E.**, **Athanasiadis, I.N.**, **Chaves, I.** ***et al.*** (2020) Enabling reusability of plant phenomic datasets with MIAPPE 1.1. *The New Phytologist*, **227**, 260–273. Available from: https://doi.org/10.1111/nph.16544

Parkinson, H., **Kapushesky, M.**, **Shojatalab, M.**, **Abeygunawardena, N.**, **Coulson, R.**, **Farne, A.** ***et al.*** (2007) ArrayExpress—a public database of microarray experiments and gene expression profiles. *Nucleic Acids Research*, **35**, D747–D750. Available from: https://doi.org/10.1093/nar/gkl995

Paskin, N. (2009) Digital Object Identifier (DOI®) System. In: Bates, M.J. & Maack, M.N. (Eds.) *Encyclopedia of library and information sciences*, Third edition. Boca Raton, FL, USA: CRC Press, pp. 1586–1592. Available from: https://doi.org/10.1081/E-ELIS3-120044418

Perkel, J. (2016) Democratic databases: science on GitHub. *Nature*, **538**, 127–128. Available from: https://doi.org/10.1038/538127a

Pommier, C., **Coppens, F.**, **Ćwiek-Kupczyńska, H.**, **Faria, D.**, **Beier, S.**, **Miguel, C.** ***et al.*** (2023) Plant science data integration, from building community standards to defining a consistent data lifecycle. In: Williamson, H.F. & Leonelli, S. (Eds.) *Towards responsible plant data linkage: data challenges for agricultural research and development*. Cham: Springer International Publishing, pp. 149–160. Available from: https://doi.org/10.1007/978-3-031-13276-6_8

Proost, S. & **Mutwil, M.** (2018) CoNekT: an open-source framework for comparative genomic and transcriptomic network analyses. *Nucleic Acids Research*, **46**, W133–W140. Available from: https://doi.org/10.1093/nar/gky336

Pühler, A. (2016) German network for bioinformatics infrastructure–de.NBI. *The German Network for Bioinformatics Infrastructure*, 8–13.

Ram, K. (2013) Git can facilitate greater reproducibility and increased transparency in science. *Source Code for Biology and Medicine*, **8**, 7. Available from: https://doi.org/10.1186/1751-0473-8-7

Rettberg, N. & **Schmidt, B.** (2012) OpenAIRE - building a collaborative Open Access infrastructure for European researchers. *LIBER Quarterly: The Journal of the Association of European Research Libraries*, **22**, 160–175. Available from: https://doi.org/10.18352/lq.8110

Rhee, S.Y., **Beavis, W.**, **Berardini, T.Z.**, **Chen, G.**, **Dixon, D.**, **Doyle, A.** ***et al.*** (2003) The Arabidopsis Information Resource (TAIR): a model organism database providing a centralized, curated gateway to Arabidopsis biology, research materials and community. *Nucleic Acids Research*, **31**, 224–228. Available from: https://doi.org/10.1093/nar/gkg076

Sansone, S.-A., **Rocca-Serra, P.**, **Field, D.**, **Maguire, E.**, **Taylor, C.**, **Hofmann, O.** ***et al.*** (2012) Toward interoperable bioscience data. *Nature Genetics*, **44**, 121–126. Available from: https://doi.org/10.1038/ng.1054

Schwacke, R., **Ponce-Soto, G.Y.**, **Krause, K.**, **Bolger, A.M.**, **Arsova, B.**, **Hallab, A.** ***et al.*** (2019) MapMan4: a refined protein classification and annotation framework applicable to multi-omics data analysis. *Molecular Plant*, **12**, 879–892. Available from: https://doi.org/10.1016/j.molp.2019.01.003

Sefton, P., **Ó Carragáin, E.**, **Soiland-Reyes, S.**, **Corcho, O.**, **Garijo, D.**, **Palma, R.** ***et al.*** (2023) RO-Crate Metadata Specification 1.1.3. https://doi.org/10.5281/zenodo.7867028

Simonyan, V., **Goecks, J.** & **Mazumder, R.** (2017) Biocompute objects—a step towards evaluation and validation of biomedical scientific computations. *PDA Journal of Pharmaceutical Science and Technology*, **71**, 136–146. Available from: https://doi.org/10.5731/pdajpst.2016.006734

Singh, J. (2011) FigShare. *Journal of Pharmacology and Pharmacotherapeutics*, **2**, 138–139. Available from: https://doi.org/10.4103/0976-500X.81919

The Galaxy Community. (2022) The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2022 update. *Nucleic Acids Research*, **50**, W345–W351. Available from: https://doi.org/10.1093/nar/gkac247

Vizcaíno, J.A., **Foster, J.M.** & **Martens, L.** (2010) Proteomics data repositories: providing a safe haven for your data and acting as a springboard for further research. *Journal of Proteomics*, **73**, 2136–2146. Available from: https://doi.org/10.1016/j.jprot.2010.06.008

Vuorre, M. & **Curley, J.P.** (2018) Curating research assets: a tutorial on the Git version control system. *Advances in Methods and Practices in Psychological Science*, **1**, 219–236. Available from: https://doi.org/10.1177/2515245918754826

Wilkinson, M.D., **Dumontier, M.**, **Aalbersberg, I.J.**, **Appleton, G.**, **Axton, M.**, **Baak, A.** ***et al.*** (2016) The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, **3**, 160018. Available from: https://doi.org/10.1038/sdata.2016.18

Winter, D., **Vinegar, B.**, **Nahal, H.**, **Ammar, R.**, **Wilson, G.V.** & **Provart, N.J.** (2007) An "Electronic Fluorescent Pictograph" browser for exploring and analyzing large-scale biological data sets. *PLoS One*, **2**, e718. Available from: https://doi.org/10.1371/journal.pone.0000718

Wolstencroft, K., **Krebs, O.**, **Snoep, J.L.**, **Stanford, N.J.**, **Bacall, F.**, **Golebiewski, M.** ***et al.*** (2017) FAIRDOMHub: a repository and collaboration environment for sharing systems biology research. *Nucleic Acids Research*, **45**, D404–D407. Available from: https://doi.org/10.1093/nar/gkw1032

Zhou, N., **Siegel, Z.D.**, **Zarecor, S.**, **Lee, N.**, **Campbell, D.A.**, **Andorf, C.M.** ***et al.*** (2018) Crowdsourcing image analysis for plant phenomics to generate ground truth data for machine learning. *PLoS Computational Biology*, **14**, e1006337. Available from: https://doi.org/10.1371/journal.pcbi.1006337

Zhou, X.-R., **Beier, S.**, **Brilhaus, D.**, **Rodrigues, C.M.**, **Muehlhaus, T.**, **von Suchodoletz, D.** ***et al.*** (2023) DataPLAN: a web-based data management plan generator for the plant sciences. https://doi.org/10.1101/2023.07.07.548147

Zhu, F., **Wen, W.**, **Cheng, Y.**, **Alseekh, S.** & **Fernie, A.R.** (2023) Integrating multiomics data accelerates elucidation of plant primary and secondary metabolic pathways. *aBIOTECH*, **4**, 47–56. Available from: https://doi.org/10.1007/s42994-022-00091-4