DataPLANT – Tools and Services to structure the Data Jungle for fundamental plant researchers

Timo Mühlhaus, Dominik Brillhaus, Marcel Tschöpe, Oliver Maus, Björn Grüning, Christoph Garth, Cristina Martins Rodrigues and Dirk von Suchodoletz

The DataPLANT consortium focuses on the continuous development and improvement of mechanisms and services for collaborative research based on sharing, enrichment and crosslinking of plant-research specific (meta)data. For this purpose, the DataPLANT tool and service chain is intended to facilitate overarching collaboration and research context management, ultimately leading to a more open and cooperative handling of research data through publication. DataPLANT follows a gradual and iterative approach, ensuring the commitment and alignment of expectations of all stakeholders. This particularly emphasizes the interaction between the community and DataPLANT.

The set of tools and microservices developed and advanced in the last couple of months focused on the pre-existing digital landscape of the average plant scientist. The first important step to data management and publication is the assisted annotation of raw data sets through the Swate Workflow Annotation Tool for Excel, which integrates the required external ontologies. The selection of the relevant metadata is simplified by provisioning of metadata templates and the use of non-integrated terms is supported by the Swate OBO Updater. The ArcCommander helps with the creation of the specific folder and file structure following the concept of the Annotated Research Context. In the future, a comprehensive workflow integration and a collaborative platform for data provenance and research sharing will emerge supporting decentralized and centralized digital processes. A central DataPLANT Hub will offer an aggregation of services and knowledge, generating a searchable compendium for research in plant biology.

1 DataPLANT core motivation

In many disciplines, scientists increasingly rely on research data management (RDM) services and infrastructures that facilitate the collection, processing, exchange and archiving of research data sets. A modern, integrated RDM enables reproducible research, the linking of interdisciplinary expertise, the sharing of research for comparison and integration of different analysis results and metadata studies, taking advantage of the immense additional knowledge gained from them. DataPLANT[1] as part of the National Research

Publiziert in: Vincent Heuveline, Nina Bisheh (Hg.): E-Science-Tage 2021. Share Your Research Data. Heidelberg: heiBOOKS, 2022.
DOI: https://doi.org/10.11588/heibooks.979.c13724 (CC BY-SA 4.0)

Data Initiative (NFDI)[2] aims to generate this added value in the field of fundamental plant research. In this domain, the (molecular) principles of plant life that determine plant growth, crop yield and biomass production are investigated. The methods used for this purpose nowadays often comprise high throughput techniques e.g. *omics and imaging techniques. These generate high-dimensional data which have to 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 annotation, standardization and contextualization of the data i.e., in a metaphoric sense, a structuring of the data jungle.

The FAIR[9] and Linked Open Data[4] Principles provide an abstract guideline for RDM. Nevertheless, besides these stated best practices it is almost always left to the initiative of individual researchers to implement them, requiring significant time and resources. To address this bottleneck, we opt for a close community-integrative approach mirrored in a three-pillar structure of i) standardization, ii) personal, and iii) technical support for research groups and individual researchers[5]. By combining technical expertise in basic plant research, information and computer sciences and infrastructure specialists, DataPLANT supports plant scientists in all aspects of RDM. It strives to advance a specific community standard for fundamental plant research (meta)data and workflow annotation and provides the necessary tools to facilitate the annotation and handling of data.
Based on the expertise of computer scientists, bioinformaticians, service providers and contribution from the community, development principles were established leading to a first set of tools and workflows has been developed and made available, the elaboration of which is detailed in this paper.

2 Fundamental design principles of the DataPLANT tool chain

Developing applications and tools that support community-driven RDM exceeds beyond writing code. Design principles provide high-level guidelines and a collection of considerations to create successful applications. In DataPLANT, tool development is always motivated by community requirements conveyed by researchers e.g. through data stewards to developers. The objective in DataPLANT is to provide incremental but regular improvements of the digital processes from the very beginning of the project. This will be achieved through iterative but multiple measures allowing a fast start and the possibility of a timely feedback from the practitioners in the field. The main platforms for exchange are the DataPLANT hub for documentation and the public code repository hosted on Github[6]. Ongoing activities are overseen by both a scientific and technical board. Additionally, we continuously survey our community to allow the swift integration of it's needs into the development process. It enables us to integrate our support tools and services in the work processes of the different laboratories.

This means we acknowledge the fact that RDM still represents a considerable additional effort for scientists and is therefore essentially an ad hoc management of experimental

data. Scientists are accustomed to documenting their research in free-text documents or tables loosely organized in file and directory structure. This leads to a preference for a flexible structure to support RDM in practice and reflects the dynamic nature of research. The strong desire to have full control over the research data that originates locally enables decentralized data management and tools that meet the researchers where they are. However, the advantages of cloud-based solutions are obvious and popular when it comes to querying and processing data. We reflect the natural behavior of the researcher in our design principles and avoid creating any type of walled garden or operation lock in. Therefore, we build our tool chains on top of existing tools and standards with an additional layer of comfort for biologists. It should be possible to perform any process in our tool chain without dedicated software. This increases opportunities for others to use and improve upon our work to embrace the open software principles.

3 ARC: a data-centric integration

A major challenge in modern RDM is the scientific integration of different decentralized data- and infra-structures. The evolving nature and needs of various research communities have led to a constant increase in heterogeneity of data standards, software and hardware solutions in the past. Now, given a transformation towards an integrated, multi-provider RDM model, this change in focus has many implications for the development and application of IT systems used in RDM. Regarding interoperability, there are two orthogonal models of thought: an application-centric and a data-centric one.

According to the notion, within an application-centric model, application, software and services are the main focus of integration. This requires a well-defined exchange of all information between the interconnected components. Consequently, it is necessary to agree on the exchange format or APIs to incorporate different functionalities. The main advantage of this approach is the ability to get the most out of legacy software and services due to the large number of systems already existing. However, each application needs to employ its own data model, which depends on specific functions and tasks. Therefore, the complexity increases by the sum of all elements that developers and users need to know in order to master such a system.

The data-centric model[7, 8] is based on an architecture in which data is the primary and permanent asset and applications are interchangeable. In such an architecture, the data model precedes the implementation of any given application. At this point, services and applications are in a state of constant change to meet user requirements and experiences or functionality extensions. In the data-centric world, the data model focuses on semantics. The structure, constraints, and validation that need to be done to the data are only secondarily included. This allows for a local and independent model to support functionality and a separation of concerns regarding system design and interoperability.

In respect to RDM, it seems natural to consider the data as the center and build the DataPLANT tool chain around it. This results in the technical realization and standardized

RDM procedures being process-oriented, meaning that each tool realizes or supports the researcher in a distinct task within the RDM cycle. Consequently, this enables the desired mixed mode of application, in which both human and machine can operate processes simultaneously or asynchronously. In addition, we thereby avoid technological barriers and embrace open software and open science, respectively.

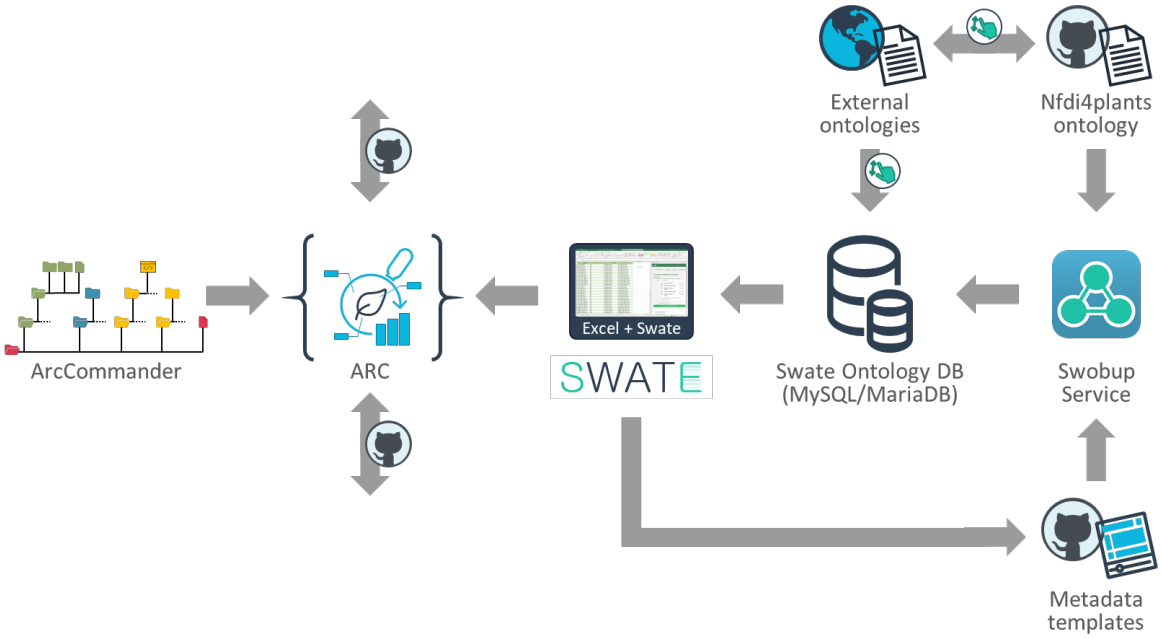

Figure 1: DataPLANT Metadata Toolchain.
The **ArcCommander** helps with the creation of the specific folder and file structure within an **ARC**. In this context, **SWATE** (Swate Workflow Annotation Tool for Excel) supports the metadata annotation process based on the **Swate DB** (Database), which integrates the required external ontologies. The selection of the relevant metadata is simplified by provision of **metadata templates** adapted to public repositories. To enable the user to use non-integrated terms, **Swobup** (Swate OBO Updater) bridges the gap to the **NFDI4plants ontology**, which stores these terms temporarily until incorporation into existing ontologies.

To realize a data centric approach for RDM in DataPLANT, we propose the Annotated Research Context[9], that captures and structures the complete research cycle meeting the FAIR requirements with low friction for the individual researcher. ARCs are self-contained and include assay/measurement data, workflow and computation results accompanied by metadata in one package. Their structure allows full user control over all metadata and facilitates usability, access, publication and sharing of the research. Thereby, ARCs are a practical implementation of existing standards encompassing the benefits of the ISA model, research object crates[10, 11] and the Common Workflow Language[12].

The ARC concept relies on a structure that partitions assays, workflow and results for a granular reuse and development. “Assays” cover biological, experimental and instrumental data including their self-contained description using the ISA model. Similarly,

“workflows” cover all digital steps of a study and contain application code, scripts and/or any other executable description of an analysis ensuring highest flexibility for the scientists. However, to ensure persistence and reproducibility, digital processes or ”workflows” comprise their own containerized running environment. The “result” data is linked to the workflows by a minimal Common Workflow Language file specifying the input and output of the process. The suggested structure for ARCs is a starting point for individual research projects and defines a framework for the organization, sharing, reuse (clone) and evolution (fork/pull request) of research projects in a fashion familiar in open-source software development.

4 Templates for convenient metadata annotation

The tools developed in DataPLANT assist in ARC creation as well as evolution through collaborating, sharing and publishing. However, the most important aspect is to increase the user engagement to actually collect metadata in a human readable and machine usable form. Therefore, a first step is to ease the publication process of research data to public repositories and shift the workload of metadata generation away from the user by template convenience.

Metadata annotation as part of the data submission routine to public endpoint repositories is often bothersome due to a high variability between repository requirements. Differences exist in both the content required (e.g. to comply with underlying metadata standards or minimal reporting guideline like MIAME[13], MINSEQE[14], MIAPPE[15]) as well as the mode format required for submission (FTP, API, webform) and also the presentation (e.g. spreadsheet, web form, check lists or mixes thereof) for meta, raw or processed data, respectively. This can become particularly inconvenient when the same metadata is submitted repeatedly or in large volumes, as in cases of individual researchers submitting the same metadata to multiple unrelated endpoint repositories or data experts requiring different formats (e.g. data champions, core facility heads) that repeatedly submit similar data to the same endpoint repository where duplication of metadata between studies is not supported.

In addition, proper metadata annotation requires use of controlled vocabularies and ontologies, which is often not intuitively supported by repository tools and can be challenging for an untrained user. Post-submission modifications and updates to datasets and metadata can be fragmented and require redundant work, e.g. when metadata on the study level needs to be updated that would affect datasets submitted to different repositories, thus eliminating version control between metadata descriptions. In summary, the wet-lab biologist can easily lose a significant amount of time adapting submission routines. Additionally, lack of flexibility -e.g. rigid requirements for metadata terms- can jeopardize the willingness or even ability to submit, if information is simply not available or the requirement is incompatible with the dataset being described.

To overcome this annotation nuisance, DataPLANT provides a growing collection of templates designed and curated by data stewards that cover the submission routine to selected end-point repositories. The template design process is initiated "backwards", starting from the requirements of end-point repositories and thereby compliance with metadata standards. Data stewards supervise the implementation of ontologies and the use of controlled vocabulary as required by the target repository, and simultaneously contribute to the development of the DataPLANT broker ontology. To provide high quality templates, data stewards cross-validate usability in an independent template reviewing process. This includes periodic mock submissions of "vanilla use-cases" through the pipeline to verify that these are fully compliant and at the same time user friendly and/or where they could be improved. High flexibility is fostered by offering a choice of modes for template distribution, use and customization.

To support this process, DataPLANT introduces SWATE[16] (Swate Workflow Annotation Tool for Excel) as a one-stop-shop (but not one-fits-all) metadata capture approach which leverages on the flexibility of the well-established ISA framework to supplement the ARC research object. As a starting point and user guide for metadata annotation. Once SWATE is installed (on the user-side) or used in the online version, templates can be loaded directly from a database. Meta information supplied by the template authors such as the target repository, study or assay type, enables the selection of suitable templates. Alternatively, SWATE templates can be easily shared and propagated via conventional routes (email, storage cloud or server), also allowing reuse of previous templates. Accordingly, the templates can be filled with or without the help of SWATE. While the latter increases the need for post-annotation curation, the former requires more expertise on the user side, but allows direct linkage to ontology references.
SWATE metadata templates are designed as a non-restrictive starting point and the user is encouraged to expand them with additional attributes. However, to minimize the need for (unsupervised) customization, data stewards interact closely with users and data type experts (champions) to integrate and align their feedback during template design. This can eventually lead to provision of very specific templates e.g. for individual core facilities or research groups to perfectly align with their daily laboratory routines. In addition to leveraging their multiplier role, this supports recording metadata at the place and time of its emergence, mitigating the need for redundant, retrospective annotation or even loss of information.

In this way, DataPLANT rethinks standardization from a purely technical towards a user-friendly, applicable perspective that aligns well with the progress of scientific innovation (e.g. new techniques and data types). With a growing user community and strong data steward interaction, the templates are continuously polished, crystallizing what information is frequently required or lost, and filling the ontology knowledge gaps. Combined with the full suite of the DataPLANT toolbox, SWATE templates lower the users' burden and workload of data publishing in the long run. However, they also allow immediate benefit through repository compliance, harmonized grammar, structure, and use of ontologies, and by guaranteeing usability independent of other DataPLANT mechanisms.

5 Swate for ontology driven metadata annotation

At first glance, the diversity of fundamental research is not ideally mirrored by the rigidity of standardization processes and requirements of metadata management. The balancing act between requirements of the researcher and standardization is especially applicable for the annotation of experimental measurement workflows. The spreadsheet-based version of the ISA standard[17] allows for ontology-driven metadata annotation of technical workflows in a simple and accessible way, compromising between free form and aligning with standardization efforts. However, finding the appropriate ontology term can be extremely tedious and often results in incomplete metadata annotation. In the DataPLANT tool chain we offer Swate to facilitate this via an integrated search function and an ontology guided metadata annotation.

Swate is an add-in that allows the user to easily annotate their data according to ISA standards. It focuses on providing an easy-to-use tool in the widely used and thus familiar environment of the most used spreadsheet editor. In order to directly incorporate collaborative mechanisms, the tool is implemented both as a modern web-based online application and as a desktop application. Fully integrated in Microsoft Excel, users can add and delete columns with specialized headers describing their data in a clear representation. By design, Swate facilitates ontology-driven development of data annotation schemes by the domain experts performing the actual research data generation. Swate features a search function for ontological terms, facilitating an ontology-driven annotation of the data. It can insert ISA-conform protocols and processes that support the DataPLANT template mechanism. By making the trade-off between free form and alignment with a standardization, e.g., ISA syntax, we believe to encourage more researchers to increase their annotation data input. Leveraging standard spreadsheet features, such as color coding, style, and markup as free comment or highlight functionality, ultimately increases the user acceptance and user experience dramatically without polluting the actual metadata information separately stored in the specialized xml dialect named spreadsheetML.

Finally, Swate simplifies mapping between models and their semantic representations in the form of the ISA model, facilitating machine readability of user-annotated data in the result.

6 Swobup and SwateDB, a team for metadata broker to bridge the ontology gap

One of DataPLANT´s core responsibilities is to reduce the effort for users as well as to increase comfort in providing human- and machine-readable metadata. Therefore, the use of controlled vocabularies is indispensable. Controlled vocabularies enable scientists to easily classify, find and reuse data. Reuse is further supported by hierarchical or relational conceptualization in form of an ontology, rendering data and metadata readable to humans and machines. The ambition of DataPLANT lies in addressing all plant research data from

the greenhouse to the lab bench to the endpoint repository. There are excellent ontology portals or Ontology Lookup Services available. These tools and services provide an easy access to all or most available ontologies. However, the problem for our community centric approach is that these services may offer similar terms from different sources leaving the user alone with a decision. This leads to a clatter in ontology references or in the worst case to ambiguity that in itself defeat the purpose to use an ontology in the first place.

In DataPLANT, Data stewards supervise the ontology selection to fine-balance between a limited scope (i.e. pre-selection of ontologies) and the flexibility of the user to provide mandatory and meaningful terminologies for data descriptions. Consequently, ontologies are selected and imported into a database called SwateDB. SwateDB comprises a controlled and practical mix of ontologies developed for specialized plant sciences topics, as well as technical terminologies required for the acquisition (omics) and analysis (bioinformatics) of biological data. Additionally, the SwateDB is the central storage for metadata annotation templates that can be managed and consumed by our tool chain.

However, experience has taught us that missing or unsuitable ontology terms and relations lead to a setback of user motivation and participation. DataPLANT provides a dedicated NFDI4plants ontology (to act as a broker and) bridge between the individual researcher and main ontology provider. This ontology enables the collection of missing vocabulary for immediate use and is also stored in the SwateDB. The automatic process handler Swobup[18] (Swate OBO Updater) simplifies this ingestion process of adding or removing (one or more) terms from an ontology and synchronizes ontology terms in OBO format and publicly hosted metadata templates either by pull request mechanism or a group of authorized users. File versioning and adaptations to the database are outsourced to a shared repository. Any change can be reverted using the repository's built-in features.

Swobup recognizes these changes and reverts to previous versions based on the principles described above. Swobup parses the OBO or template file and incorporates the changes into the SwateDB database. In order to work with Swobup a webhook has to be defined in Github and configured, that it sends a SSL encrypted HTTP Post request to a previously configured URL every time files are changed in the repository. This process allows an immediate update process including version control and history driven by the community. Therefore, anyone is directly or indirectly (via pull request) able to update templates or ontologies without delay and a minimum amount of guidance. This process enables DataPLANT to act as an ontology broker, collecting required terms from the community in the NFDI4plants ontology and forwarding them to the main ontology provider.

7 ARC Commander – Support Tool

For the community of plant researcher, experimental metadata in a structured user-friendly format are most useful to reuse research data and generate new biological knowledge. However, it is advantageous to argument these data with supplementary organizational metadata. Additionally, a solution to organize and manage metadata and research

data practically with low friction for the user needs to be provided. Therefore, DataPLANT introduced the ARC into the research data management landscape. The ARC is an intuitive specification for the primary setup of an experiment and collaboration environment for the storage of research data including context like metadata. Most importantly, the ARC layout follows a specific file and folder structure derived from the RO crate standard and components are registered in a central investigation file that follows the ISA model specification.

Although these requirements are minimal, assisting the researcher in these repetitive tasks and providing structured guidance is beneficial and reduces friction and workload on the side of the user. This is the main aim of the tool ArcCommander[19]. Essentially, this tool provides automation and assistance with processes following the ARC specification. The ArcCommander can be executed to initialize an ARC, creating the basic folder structure, and setting the working environment. Additionally, it can be used to create and modify sub-branches of the ARC, such as assays and workflows. By using the ArcCommander, the researchers are guided during the process and can create or maintain the ARC without needing explicit knowledge of the ARC structure. Besides ARC specifics, general naming recommendations shared across operating systems are adhered to by the ArcCommander. Following the ARC or ISA model respectively, a central registry, called investigation, is stored as a file in which all components of the ARC are registered. Manual registry synchronization upon addition of further content would be time-consuming and error-prone, but can be achieved automatically using the ArcCommander.

In its current state, the ArcCommander is implemented as a command line tool. Often there are experiments that are very similar to each other in some characteristics. For example, proteomics measurements performed in the same laboratory might follow the same protocols. In this case, the resulting ARCs are also likely to have some properties in common. Here, repetition can be easily reduced by concatenating the commands using a script. Commands and parameters are designed to automatically guide the researcher through the process of creating an ARC. This guidance is realized by providing a hierarchically structured and extensively labeled command set, which can be easily and purposefully browsed for the command of interest. Additionally, the user experience is enhanced by a text editor enabling to automatically generate metadata schemata. Instead of specifying all arguments in the command line, a text-based form is created and presented to the user that handles the metadata retrieval.

Enabling successful data sharing, working in teams and information exchange between researchers are the fundamental tasks in RDM. The ArcCommander supports collaborative work by leveraging Git-based version control to keep track of file change history as well as user interactions and contributions. The ArcCommander implements a convenience layer on top of Git to enable synchronization functionality for non-expert users. Besides using standard Git, it can handle large files which are common in research using Git-LFS (Large File Storage). By this, researchers can easily share and control their state of the ARC without additional efforts.

The modularity of the ArcCommander adequately accounts for the dynamic design principles and flexible extensibility required to maintain and extend ARC functionality. A process that requires the most extensibility might be the data publication based on ARCs. After seamless creation of the ARC, a major interest for researchers is the distribution of their research data. An increasing need has been to publish data in a centralized repository that prescribes individual technical metadata information and format requirements. Meeting these requirements imposes a significant additional burden on the researcher. Therefore, the ArcCommander aims to provide automatic export to different central data repositories and to support transformation of the ARC along different formats and requirements. Already included in the ArcCommander is the possibility to easily export the metadata available in "ISAxlsx" format to "ISAJson" and "ISATab" standard formats. This allows for seamless interoperability with available ISA Tools and all central repositories complying with standard ISA model specifications. For the future, a successive extension of the export functionality is planned in order to provide compatibility to additional data repositories and community resources.

8 Workflow and data integration with the Galaxy Gateway

Due to our data-centric approach in DataPLANT, all digital processes are centered around the ARC. This includes or especially applies to data processing and analysis workflows. Galaxy[22] is an integral part of DataPLANT regarding workflow management. Dedicated tools for the plant science community will be integrated during the DataPLANT project and will extend the portfolio already available at Galaxy for Plants[21]. For years, Galaxy has made advanced bioinformatics software accessible to scientists worldwide by providing an intuitive web interface to these applications while fostering reproducibility through the automatic creation of re-runnable protocols of each analysis. The Galaxy community is one of the largest bioinformatics communities world-wide[22]. It provides over 7000 tools and a plenitude of bioinformatics and data processing workflows useful for researchers from the fundamental plant research community. The core framework offers various abstraction layers that offer various extension points and adopt Galaxy to new technologies, while keeping the system maintainable since 16 years.

Tools in Galaxy are independent of each other and contain rich metadata annotations, including all their dependencies. Those dependencies are resolved via different Galaxy plugins for Modules, Conda, Docker, Singularity or others. For truly reproducible research the Galaxy community recommends different approaches depending on the degree of reproducibility and cost. Conda for more flexible and cost efficient tool dependency management and containers for elaborated and isolated environments that are more cost intensive in maintenance. For both scenarios the Galaxy community offers solutions with Bioconda[23] and BioContainers[24].

Another subsystem in Galaxy is the handling of user data. Galaxy supports different kinds of data storages, ranging from hierarchical POSIX storages, to S3 or iRODS. Those can be bound to users, groups or roles and enable flexible quota assignments per object store.

A similar system allows users to browse and import public data deposited on S3, SFTP, Webdav or Dropbox accounts. Galaxy also supports the export of research artefacts, like workflow invocations. Currently, those can be exported as BioComputeObjects and support for ResearchObjects and ARCs are planned.

9 Conclusion and Outlook

Dedicated to structuring the data jungle for fundamental plant researchers, the DataPLANT consortium started its operation by assembling a suite of tools that should greatly facilitate efforts of proper research data management. As a simple structural scaffold, the Annotated Research Context is introduced, which intrinsically follows FAIR requirements.

As a starting point, the ArcCommander simplifies the generation of the ARC folder and file structure. Following the ISA model, this includes a central registry in the form of an automatically updated investigation file. For researchers the annotation of metadata seems to be the most tedious task of the RDM cycle. Swate supports the user during this process. SWATE offers a set of ontologies and metadata templates pre-selected and curated by data stewards and facilitates simple re-use of metadata. At the backend, this is enabled by Swopub, which continuously synchronizes SWATE with the SwateDB according to adaptations to templates and the NFDI4plants broker ontology. As a result of user feedback, it was already possible to create a set of initial templates that converged the user needs with the requirements of corresponding endpoint repositories. This is currently expanded to cover centralized computer-based workflows and integrate already available services such as Galaxy or nf-core[25].

According to DataPLANT's prevailing data-centric view, the listed tools serve to improve user-friendliness based on the current state of the art. Everybody from the open source community is encouraged to take the initiative and adapt existing or own tools to the changing needs or to inquire us directly. Due to the modular structure and ARC centric integrative approach in DataPLANT a continuous improvement of our services comes naturally. DataPLANT envisions a cloud version of the tool chain to be integrated centrally in the DataPLANT Hub in the future. Besides providing a public website that gives information about the project, shares news, and provides links to the project's social media channels and Git repositories, the DataPLANT Hub will create a central environment for the community. A key component will be the integrated search and exploration engine for research data using annotated metadata. Thus, the DataPLANT Hub will be a key component to make research data FAIR.
Thus, an orderly floral bouquet of user-oriented tools can blossom out of the overgrown jungle.

Acknowledgements

We acknowledge support for DataPLANT 442077441 through the German National Research Data Initiative (NFDI 7/1), the TTRR 175 (INF project), and CEPLAS is supported by Deutsche Forschungsgemeinschaft within the Excellence Initiative (EXC 1028) and under Germany's Excellence Strategy – EXC 2048/1 – project 390686111.

ORCID IDs

- Timo Mühlhaus https://orcid.org/0000-0003-3925-6778
- Dominik Brillhaus https://orcid.org/0000-0001-9021-3197
- Marcel Tschöpe https://orcid.org/0000-0002-3731-7664
- H. Lukas Weil https://orcid.org/0000-0003-1945-6342
- Oliver Maus https://orcid.org/0000-0002-8241-5300
- Björn Grüning https://orcid.org/0000-0002-3079-6586
- Christoph Garth https://orcid.org/0000-0003-1669-8549
- Cristina Martins Rodrigues https://orcid.org/0000-0002-4849-1537
- Dirk von Suchodoletz https://orcid.org/0000-0002-4382-5104

Bibliography

[1] DataPLANT. https://www.nfdi4plants.de/ (accessed: 11. May 2021).

[2] NFDI. https://www.nfdi.de/ (accessed: 11. May 2021).

[3] Wilkinson, M. D., M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, et al. 2016. The fair guiding principles for scientific data management and stewardship. Scientific Data 3, Nr. 1. https://doi.org/10.1038/sdata.2016.18

[4] Bizer, C., T. Heath, K. Idehen and T. Berners-Lee. 2008. Linked data on the web (LDOW2008). Proceeding of the 17th international conference on World Wide Web – WWW '08. doi:https://doi.org/10.1145/1367497.1367760

[5] von Suchodoletz, D. , T. Mühlhaus, J. Krüger, B. Usadel, C. Martins Rodrigues, "DataPLANT – Ein NFDI-Konsortium der Pflanzen-Grundlagenforschung", Bausteine Forschungsdatenmanagement, no. 2. German:46-56 (2021). https://doi.org/10.17192/bfdm.2021.2.8335

[6] DataPLANT - Services and infrastructures to support FAIR Data science and good data management practices within the plant basic research community. GitHub. https://github.com/nfdi4plants (accessed: 11. May 2021).

[7] Liu, Z., S. Fan, H. Jiannan Wang and J. L. Zhao. 2017. Enabling effective workflow model reuse: A data-centric approach. Decision Support Systems 93: 11–-25. doi:https://doi.org/10.1016/j.dss.2016.09.002

[8] Wuyts, R., S. Ducasse and O. Nierstrasz. 2005. A data-centric approach to composing embedded, real-time software components. Journal of Systems and Software 74, Nr. 1: 25–-34. doi:https://doi.org/10.1016/j.jss.2003.05.004

[9] ARC. GitHub. https://github.com/nfdi4plants/ARC (accessed: 11. May 2021).

[10] Carragáin, E. Ó., C. Goble, P. Sefton and S. Soiland-Reyes. 2019. A lightweight approach to research object data packaging. Zenodo. doi: https://doi.org/10.5281/zenodo.3250687

[11] eScience Lab at The University of Manchester. Research Object Crate. researchobject.org. https://www.researchobject.org/ (accessed: 11. May 2021).

[12] Amstutz, P., M. R Crusoe, N. Tijanić, B. Chapman, J. Chilton, M. Heuer, A. Kartashov, et al. 2017. Common Workflow Language, v1.0. eScholarship, University of California. 1. April. https://escholarship.org/uc/item/25z538jj (accessed: 11. May 2021).

[13] Brazma, A., P. Hingamp, J. Quackenbush, G. Sherlock, P. Spellman, C. Stoeckert, J. Aach, et al. 2001. Minimum information about a microarray experiment (MIAME)—toward standards for microarray data. Nature Genetics 29, Nr. 4: 365–-371. doi:https://doi.org/10.1038/ng1201-365

[14] Dimitrova, M., R. Meyer, P. L. Buttigieg, T. Georgiev, G. Zhelezov, S. Demirov, V. Smith and L. Penev. 2020. A Streamlined Workflow for Conversion, Peer-Review and Publication of Omics Metadata as Omics Data Papers. doi:https://doi.org/10.20944/preprints202009.0357.v1

[15] Papoutsoglou, E. A., D. Faria, D. Arend, E. Arnaud, I.s N. Athanasiadis, I. Chaves, F. Coppens, et al. 2020. Enabling reusability of plant phenomic datasets with MIAPPE 1.1. New Phytologist 227, Nr. 1: 260–-273. doi: https://doi.org/10.1111/nph.16544

[16] Swate – Excel Add-In for annotation of experimental data and computational workflows. GitHub.https://github.com/nfdi4plants/Swate (accessed: 11. May 2021).

[17] Sansone, S.-A., P. Rocca-Serra, A. Gonzalez-Beltran, D. Johnson and ISA Community. 2016. ISA Model and Serialization Specifications 1.0. Zenodo. October 28, 2016. doi:https://doi.org/10.5281/zenodo.163640

[18] Swobup - Swate DB update tool. GitHub. https://github.com/nfdi4plants/Swobup (accessed: 11. May 2021).

[19] ArcCommander - Tool to manage your ARCs. GitHub. https://github.com/nfdi4plants/ArcCommander (accessed: 11. May 2021).

[20] Afgan, E., D. Baker, B. Batut, M. van den Beek, D. Bouvier, M. Čech, J. Chilton, et al. 2018. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. Nucleic Acids Research 46, Nr. W1. doi:https://doi.org/10.1093/nar/gky379

[21] Galaxy for Plant Biology. Galaxy. https://plants.usegalaxy.eu/ (accessed: 14. May 2021).

[22] An open source Git extension for versioning large files. Git Large File Storage.https://git-lfs.github.com/ (accessed: 11. May 2021).

[23] Grüning, B., R. Dale, A. Sjödin, B. A. Chapman, J. Rowe, C. H. Tomkins-Tinch, R. Valieris and J. Köster. 2018. Bioconda: sustainable and comprehensive software distribution for the life sciences. Nature Methods 15, Nr. 7: 475–-476. doi: https://doi.org/10.1038/s41592-018-0046-7

[24] da Veiga Leprevost, F., B. A Grüning, S. Alves Aflitos, H. L Röst, J. Uszkoreit, H. Barsnes, M. Vaudel, et al. 2017. BioContainers: an open-source and community-driven framework for software standardization. Bioinformatics 33, Nr. 16: 2580–-2582. doi: https://doi.org/10.1093/bioinformatics/btx192

[25] Ewels, Phil. nf-core: A community effort to collect a curated set of analysis pipelines built using Nextflow. nf-core. https://nf-co.re/ (accessed: 14. May 2021).