DataPLANT – Harnessing the Power of Ontologies for FAIR Research Data Management

Kathryn Dumschott[1], Hannah Dörpholz[1], Kevin Frey[2], Marcel Tschöpe[3], Heinrich Lukas Weil[2], Timo Mühlhaus[2], Dirk von Suchodoletz[3], Björn Usadel[14], Angela Kranz[1]

[1]IBG-4 Bioinformatics, BioSC, Forschungszentrum Jülich;
[2]Computational Systems Biology, RPTU University of Kaiserslautern;
[3]Computer Center, University of Freiburg, Freiburg im Breisgau;
[4]Institute for Biological Data Science, CEPLAS, Heinrich Heine University, Düsseldorf

The NFDI funded DataPLANT consortium aims to provide a sustainable and user-friendly data management platform for the fundamental plant research community. DataPLANT has developed tools and services that encourage open and collaborative research, facilitate the annotation of metadata, and unify the use of ontologies. DataPLANT aims to establish a foundation that enables scientists to effortlessly use and access specific ontologies as well as to expand ontologies with missing terms in order to increase the FAIRness of their research data.

The center of DataPLANT´s developments is the Annotated Research Context (ARC), a data-centric approach to capturing and structuring the entire research cycle. As a structural ontology, the ARC container ontology is designed to help researchers contextualize their data within the ARC and easily compare it to other public ARCs, facilitating the linking of already acquired information to gain knowledge and answer new research questions. Metadata annotation within ARCs is supported by the Swate tool. Swate is linked to the Swate database (SwateDB), which stores a collection of ontologies to facilitate standardized metadata annotation. This collection includes a selection of established ontologies as well as the DataPLANT biology ontology (DPBO), a "broker ontology', which contains missing terms that do not yet appear in known, established ontologies.

1 Introduction

In recent decades, proper research data management (RDM) has become increasingly important with the advent of high throughput methods such as omics (transcriptomics, proteomics, etc.) and imaging. While these methods are incredibly useful for how much information they can provide a researcher, the sheer volume of data produced

Publiziert in: Vincent Heuveline, Nina Bisheh und Philipp Kling (Hg.): E-Science-Tage 2023. Empower Your Research – Preserve Your Data. Heidelberg: heiBOOKS, 2023. DOI: https://doi.org/10.11588/heibooks.1288.c18067 (CC BY-SA 4.0)

requires additional support and infrastructure to facilitate the correct collection, processing and storage of the data. Additionally, data must be properly integrated before any meaningful interpretation can take place. For this reason, the DataPLANT DataHUB (Bauer et al. 2023), an RDM platform that enables the proper storage and annotation of data, increasing its reusability, is important to the fundamental plant science community. As part of the National Research Data Initiative (NFDI; Hartl, Wössner, and Sure-Vetter 2021; Verein Nationale Forschungsdateninfrastruktur (NFDI) e.V. 2023), the DataPLANT (Martins Rodrigues et al. 2021; DataPLANT Consortium 2023) consortium aims to support plant scientists in managing their research data, including data organization, storage and metadata annotation, while adhering to the FAIR principles (Wilkinson et al. 2016). To do so, DataPLANT has developed tools and services that work together seamlessly to store data and annotate metadata quickly and efficiently. Git based versioning (Git community 2023) is employed within the DataPLANT DataHUB to provide complete transparency and version control of all tools, thereby encouraging community contribution and engagement. To encourage the reusability and interoperability of data, ontologies are incorporated into the platform in two ways: as a structural ontology and as an ontology service (Figure 1). By harnessing the potential of ontologies, DataPLANT is able to improve data standardization and metadata annotation, all the while facilitating the linking of previously acquired data for novel discoveries.

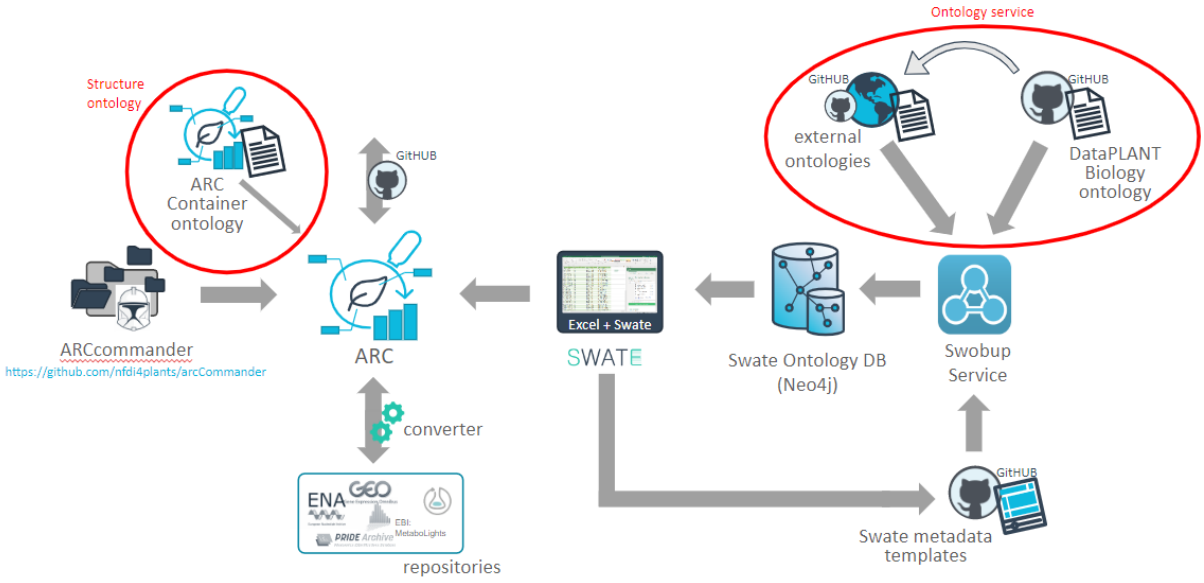

Figure 1: Outline of the DataPLANT DataHUB tools and services. Ontologies are incorporated in two main locations to facilitate FAIR plant research data.

In recent years, ontologies have become important tools for the standardization of data annotation and data reusability in the plant sciences, promoting FAIR principles. By providing unique identifiers for concepts within a domain and describing relationships between them, ontologies ensure that data is structured and can be easily understood by both humans and machines, allowing for machine-based reasoning (Walls et al. 2012). The recorded relationships between terms allows data to not only be machine-readable, but

also facilitates data to be processed in a biologically relevant way, which in turn enables the integration of different data sets.

The DataPLANT consortium harnesses this unique power of ontologies to assist plant scientists in managing their research data in a sustainable and FAIR way. While the platform described here is currently focused on plant research data, the concept can be easily adjusted to meet the requirements and needs of other scientific disciplines. DataPLANT encourages FAIR RDM for all and is therefore actively promoting the presented ontology concept to other scientific communities.

2 The ARC container ontology

FAIR digital objects (FDO) are at the core of all considerations and developments in DataPLANT. To implement a strict data centric approach for RDM, the Annotated Research Context (AC; see Garth et al. 2022; NFDI4Plants 2023a) was designed to capture and structure the complete research cycle to meet the FAIR requirements with low friction for the individual researcher in plant biology. ARCs are self-contained structures that include all biological, measurement, and computational data, as well as relevant metadata, produced during a scientific investigation. Components of ISA (investigation, study and assay; ISA Tools 2023), as well as the CWL (common workflow language; CWL Project 2023) runs and workflows are incorporated to increase the shareability of data. Currently under development, the ARC container ontology (NFDI4Plants 2023b) represents the metadata structure of an ARC (Figure 2). Within the ARC container ontology, the three sub-categories "investigation", "study" and "assay" are represented as a hierarchical structure, with each branch incorporating the required metadata defined by the ISA model. These include "ontology sources' under investigation, "study protocols and materials" under study and "assay technologies and data files" under assay. Major ISA concepts such as "person", "publication" and "ontology source" are represented as classes, while attributes relating to each class are represented as data properties. The individual classes are connected through object properties, creating semantic context and giving the ontology its structure.

The ARC container ontology is important for information inference, or the ability to integrate data from multiple sources. This facilitates the identification of connections and patterns that might not be apparent when investigating individual datasets. This, in turn, enables the linking of already acquired information to gain novel insights and answer new research questions. For instance, if two experiments use varying methods to measure the same set of variables, the ARC container ontology can help to align the data and make meaningful comparisons between them. Another advantage of using the ARC container ontology for information inference is the ability to apply reasoning and inference algorithms to the data. This approach can reveal implicit relationships and dependencies that are not explicitly stated in the metadata, leading to a deeper understanding of the data.

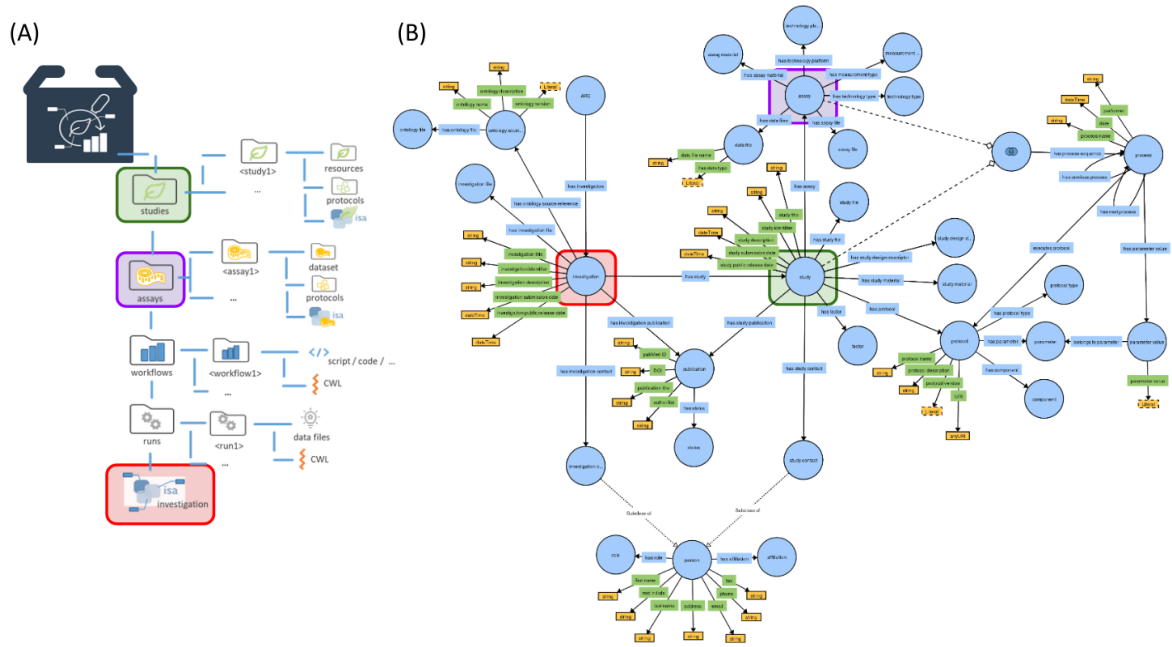

Figure 2: (A) The Annotated Research Context (ARC) is the central component of the DataPLANT DataHUB, where research data is structured and stored. (B) Excerpt of the ARC container ontology, which represents the metadata structure of the ARC. Classes are shown in blue circles, object properties are shown in blue rectangles, datatype properties and their datatype ranges are shown in green and yellow rectangles, respectively. The ontology was visualized using WebVOWL (`https://service.tib.eu/webvowl`).

3 The DataPLANT ontology service

A crucial component of RDM is the complete and proper annotation of metadata. Metadata, or the data about data, gives the context and background of studies and assays performed within an investigation. Raw data is essentially meaningless without accompanying metadata that provides information about the sample acquisition process and machine parameters used for measurement. For this reason, DataPLANT has developed a suite of seamlessly integrated tools for the annotation of metadata that work together to encourage plant researchers to annotate their data as early as possible in the experimental timeline (Mühlhaus et al. 2022).

Swate (NFDI4Plants 2023d), DataPLANT's own ontology-driven metadata annotation tool, simplifies the annotation of assay and study metadata for plant researchers. Swate is available both online and as a plugin for Microsoft Excel, enabling users to incorporate standard spreadsheet features into their metadata annotation protocol, making it convenient and user-friendly. Users can easily modify their ISA-compliant sheets by adding or removing building blocks in order to describe the necessary metadata in a clear and concise way. Building blocks can cover relevant characteristics, parameters, factors, com-

ponents and protocols. When a building block is added to the spreadsheet as a header, it is tagged with the appropriate ontology term, referred to as the "parent term". To facilitate data input, users can then utilize the "Ontology term search" tab to fill in the values corresponding to the building block heading. These values are also tagged with the appropriate ontology term, known as "child term" (Figure 3). This functionality allows Swate to not only tag the headings of metadata sheets with ontology terms, but to also tag the respective column values, ensuring comprehensive and structured annotation. While Swate provides pre-suggested terms for selection, users have the flexibility to add their own terms not linked to a current ontology term. This flexibility accommodates diverse research needs while still promoting the essential task of metadata annotation. To assist scientists in initiating the metadata annotation process, Swate also provides a variety of different templates which already include building blocks based on metadata required by repositories or minimum information standards (see section 4).

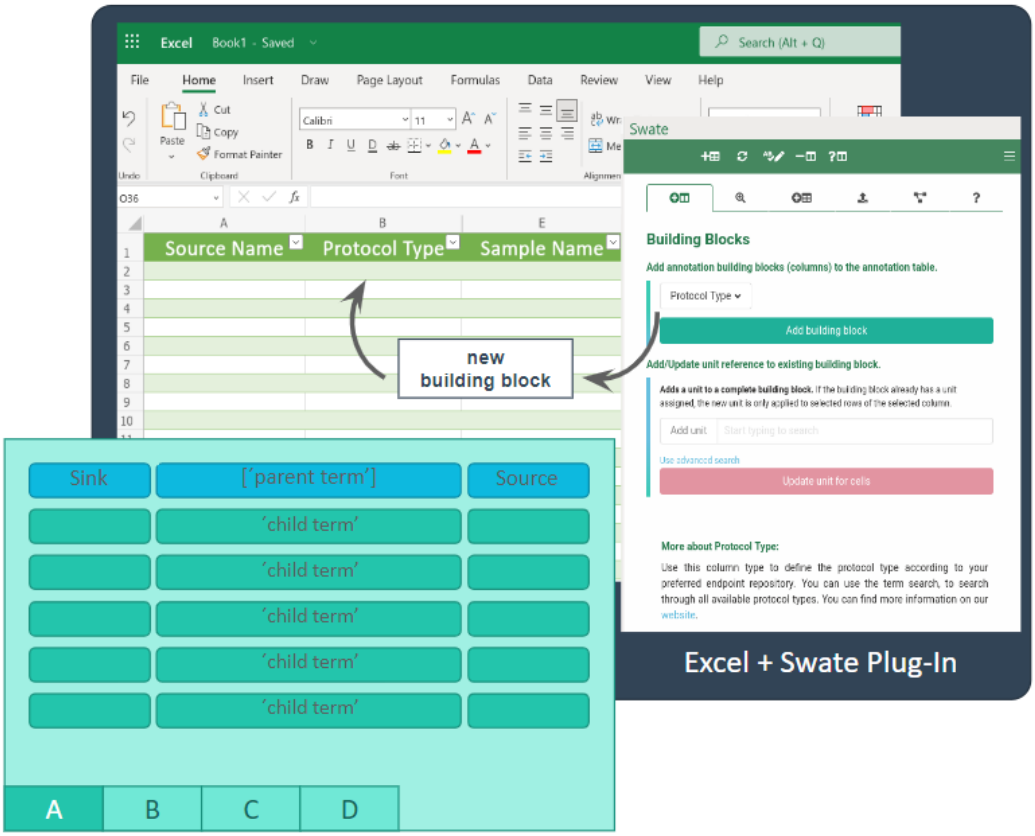

Figure 3: Swate incorporates ontologies within the SwateDB to facilitate the annotation of study and assay metadata.

To adequately describe the metadata essential for plant science experiments and investigations, a diverse range of ontologies spanning various domains is required. The DataPLANT ontology team collaborates closely with the plant science community to identify ontologies that encompass a substantial number of terms required for metadata annotation of plant-related experiments. Selected ontologies are imported into a database

called SwateDB. Currently, included ontologies cover domains related to the plant sciences, such as the Plant Ontology (PO; see Walls et al. 2012), more general biology and chemistry domains, such as the Chemical Entities of Biological Interest (CheBI; see Degtyarenko et al. 2007), as well as technical terminologies required for the acquisition and analysis of biological data. The full list of external ontologies currently included in SwateDB can be found in the nfdi4plants_ontology Github repository (`https://github.com/nfdi4plants/nfdi4plants_ontology/blob/main/ext_ontologies.include`; see NFDI4Plants 2023c). While the included external ontologies are mostly applicable for fundamental plant research, further ontologies relevant for other scientific disciplines (e.g. microbiology) can easily be added to SwateDB and be used for metadata annotation. In addition to established ontologies, DataPLANT has developed its own DataPLANT Biology Ontology (DPBO), which is also included in SwateDB. The DPBO serves two primary goals: addressing the ontology gap by collecting missing vocabulary required to annotate metadata and acting as a middleman between the researcher and the main ontology provider. Community contributions are encouraged and suggestions for new terms or improvements to already existing terms can be easily submitted via the DataPLANT helpdesk[1] or the Issues tab located at the nfdi4plants_ontology repository[2]. Once a term suggestion or improvement has been submitted, the ontology team incorporates the information into the OBO file (Figure 4). This process involves assigning a unique ID to the term, adding information such as the term definition or relevant synonyms, and incorporating the term into the ontology structure (for example, via is_a to delineate parent-child relations). As terms are suggested by the fundamental plant research community, the DataPLANT ontology team actively liaises with the main ontology providers, serving as a broker between the researchers and the ontology providers.

Once the OBO file has been saved, the term undergoes an automatic process facilitated by the Swate OBO Updater (Swobup; see NFDI4Plants 2023f), enabling its integration into SwateDB and subsequent availability within Swate (Figure 5). During this automatic step, Swobup reads OWL-compatible files retrieved in an update in a Git repository (Git community 2023), and applies the changes to the graph database (Neo4j[3]). With Git, the ontology source files are tracked, ensuring the integrity and versioning of the files. This approach enables community engagement in ontology development through standard mechanisms successfully employed in open source software development. Consequently, ontology changes and updates can be swiftly incorporated into SwateDB, making them readily available in Swate. This includes changes to the DPBO as well as the collection of external ontologies, although, in contrast to the external ontologies, terms deleted from the DPBO will also be removed from SwateDB. The file versioning feature further facilitates easy reversal to previous file versions, which can seamlessly be incorporated into Swate via Swobup.

1 `https://helpdesk.nfdi4plants.org`
2 `https://github.com/nfdi4plants/nfdi4plants_ontology`
3 `https://neo4j.com`

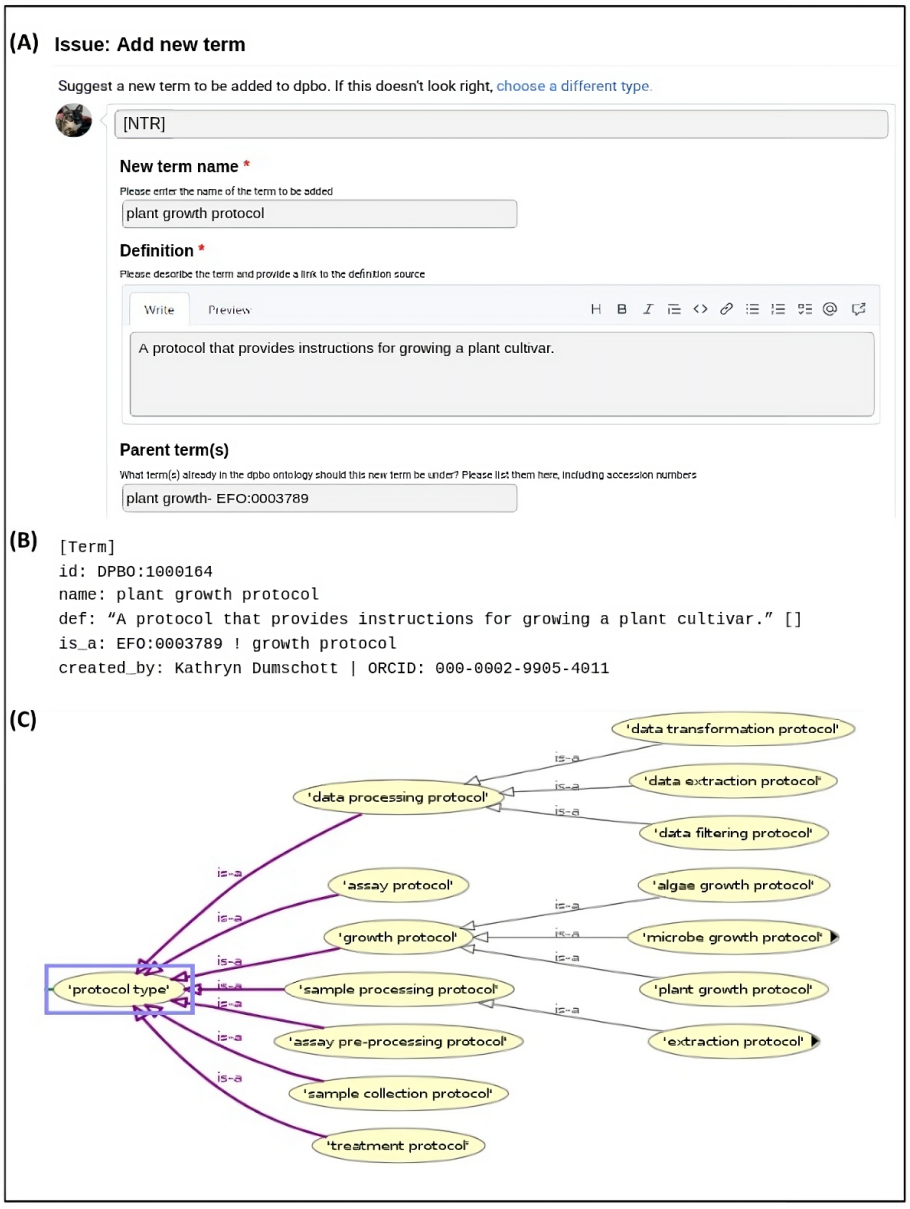

Figure 4: The DPBO is curated with the help of the scientific community (A) users can submit suggestions for new ontology terms or additions to already existing terms via the Issues tab at `https://github.com/nfdi4plants/nfdi4plants_ontology` or the DataPLANT Helpdesk (`https://helpdesk.nfdi4plants.org`) (B) the DataPLANT ontology team incorporates the term into DPBO (C) once the DPBO file has been updated with the new term, it is written to SwateDB by Swobup, making the term readily available within Swate.

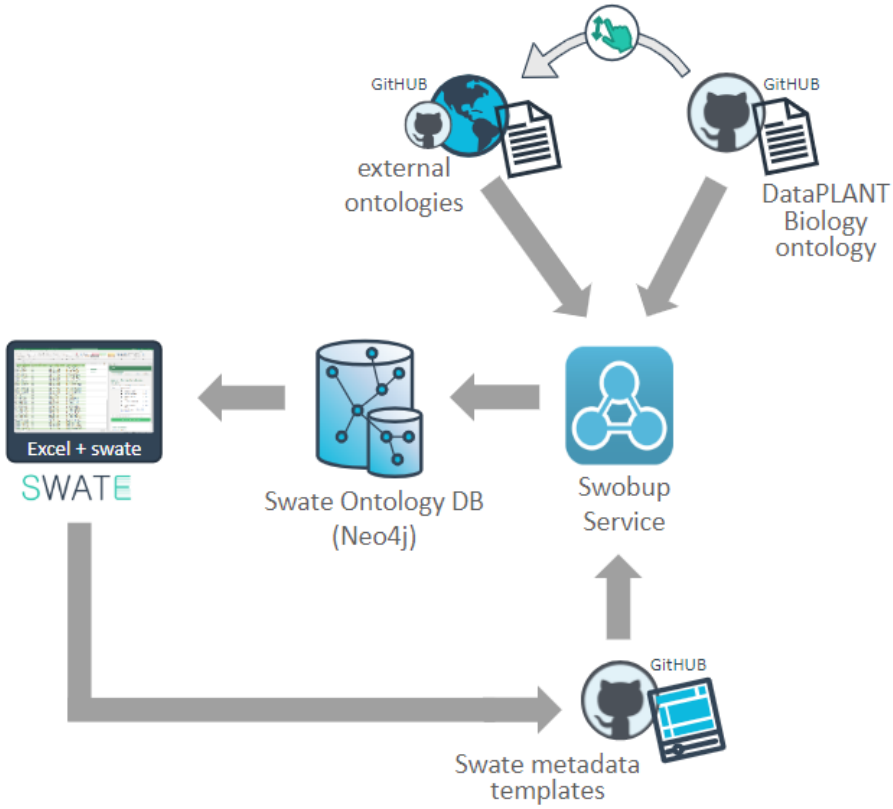

Figure 5: The Swobup service and SwateDB work together to make the annotation of metadata sheets within Swate possible.

4 Templates

A crucial first step in FAIR RDM is knowing what metadata should be included when annotating a study or assay, before even knowing what ontologies are relevant. Necessary metadata can be found in checklists required by repositories when uploading data, or else based on minimum information standards such as MIAPPE (Papoutsoglou et al. 2020), MIAME (Brazma et al. 2001), or MIMARKS (Yilmaz et al. 2011). To assist researchers in beginning the process of metadata annotation, Swate provides a collection of templates, which outline the metadata required by the repository or minimum information standard, located under the "Templates tab" in Swate. Researchers can search for relevant templates or filter templates by tags. When a template is selected, the researcher is returned to the main page of Swate, where all the building blocks contained in the template are displayed. The researcher can then click "Add template" to add the template building block to the Excel sheet. Selected templates can be used as-is or else amended to fit the exact need of the researcher, so only building blocks not yet included in the table will be added, and any unnecessary building blocks can be easily removed. To ensure total control and flexibility over the metadata annotation process, users also have the option of creating and saving their own metadata templates. These can reflect the specific experiments or

protocol performed during an investigation and can be shared with other members in the group to help standardize metadata annotation within institutes and to encourage new members to do so as early in the investigation timeline as possible. A template is created in the same way as Swate metadata sheets (described above), but with an additional SwateMetadataSheet, which must be filled out to create a function template. Once the template has been added to the repository, it is incorporated into SwateDB via the same Swobup process described above. A detailed description of how to create customized templates can be found in the Swate template Github repository (NFDI4Plants 2023e). As with all of DataPLANT's tools and standards, the use of Git when creating and uploading templates ensures data integrity and versioning control.

5 Conclusions

With the DataPLANT DataHUB, the DataPLANT consortium aims to provide a sustainable and well-annotated data management platform that encourages open and collaborative research and facilitates annotation of metadata. The consortium has developed flexible, adaptable tools and services that incorporate ontologies in two ways to support researchers in increasing the FAIRness of their plant-related research data.

Firstly, the ARC Container ontology is an organizational representation of the data-centric Annotated Research Context (ARC). While still under development, the ARC ontology already incorporates components of ISA (investigation, study and assay) and is being actively expanded to conclude CWL runs and workflows. The ontology plays a crucial role for information inference and enables the linking of acquired data to gain novel insights that may not be apparent when investigating individual datasets.

Furthermore, ontologies are incorporated within the ontology service, which supports the annotation of metadata sheets via the Swate and Swobup tools. The DBPO acts as a broker ontology, giving researchers the opportunity to efficiently add new terms that may be absent from established ontologies. The DataPLANT ontology team then acts as the middle man, curating the terms and suggesting them back to the relevant established ontologies. In addition to ontologies, a series of templates is provided to aid researchers in beginning their journey into metadata annotation. These are based on checklists included in repositories and minimum information standards. Researchers can select available templates, subsections of available templates, or else design their own based on experiments or protocols commonly performed during their research.

Most importantly, the tools and services within the platform are data centric and serve to improve the user experience while implementing current state of the art practices and versioning control. The flexibility of the DataPLANT concept means it can be easily adapted to fit the needs of researchers in other scientific domains. In the spirit of open source, DataPLANT encourages users to openly contribute and collaborate, thereby continuously improving the RDM landscape for all.

Acknowledgements

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

References

Bauer, Jonathan, Marcel Tschöpe, Julian Weidhase, Timo Mühlhaus, Christoph Garth, Gajendra Doniparthi, Holger Gauza, Louisa Perelo, Cristina Martins Rodrigues, and Dirk von Suchodoletz. 2023. "From DataPLANT's DataHUB to DataPUB(lication)". In *International Workshop on Science Gateways.* Accepted for publication.

Brazma, Alvis, Pascal Hingamp, John Quackenbush, Gavin Sherlock, Paul Spellman, Chris Stoeckert, John Aach, et al. 2001. "Minimum information about a microarray experiment (MIAME)—toward standards for microarray data". *Nature Genetics* 29 (4): 365–371. DOI: https://doi.org/10.1038/ng1201-365.

CWL Project. 2023. "Common Workflow Language". Visited on May 30, 2023. https://www.commonwl.org/.

DataPLANT Consortium. 2023. "DataPLANT". Visited on May 30, 2023. https://www.nfdi4plants.de/.

Degtyarenko, Kirill, Paula de Matos, Marcus Ennis, Janna Hastings, Martin Zbinden, Alan McNaught, Rafael Alcantara, Michael Darsow, Mickaël Guedj, and Michael Ashburner. 2007. "ChEBI: a database and ontology for chemical entities of biological interest". *Nucleic Acids Research* 36 (Database): D344–D350. DOI: https://doi.org/10.1093/nar/gkm791.

Garth, Christoph, Jonas Lukasczyk, Timo Mühlhaus, Benedikt Venn, Jens Krüger, Kolja Glogowski, Cristina Martins Rodrigues, and Dirk Von Suchodoletz. 2022. "Immutable yet evolving: ARCs for permanent sharing in the research data-time continuum". In *E-Science-Tage 2021: Share Your Research Data,* edited by Vincent Heuveline and Nina Bisheh, 366–373. heiBOOKS. DOI: https://doi.org/10.11588/heibooks.979.c13751.

Git community. 2023. "Git". Visited on May 30, 2023. https://git-scm.com/.

Hartl, Nathalie, Elena Wössner, and York Sure-Vetter. 2021. "Nationale Forschungsdateninfrastruktur (NFDI)". *Informatik Spektrum* 44 (5): 370–373. DOI: https://doi.org/10.1007/s00287-021-01392-6.

ISA Tools. 2023. "ISA Model and Serialization Specifications". Visited on May 30, 2023. https://isa-specs.readthedocs.io/en/latest/isamodel.html.

Martins Rodrigues, Cristina, Dirk von Suchodoletz, Timo Mühlhaus, Jens Krüger, and Björn Usadel. 2021. "DataPLANT – Ein NFDI-Konsortium der Pflanzen-Grundlagenforschung". *Bausteine Forschungsdatenmanagement,* number 2 (2): 46–56. DOI: https://doi.org/10.17192/bfdm.2021.2.8335. https://bausteine-fdm.de/article/view/8335.

Mühlhaus, Timo, Dominik Brillhaus, Marcel Tschöpe, Oliver Maus, Björn Grüning, Christoph Garth, Cristina Martins Rodrigues, and Dirk Von Suchodoletz. 2022. "DataPLANT – Tools and Services to structure the Data Jungle for fundamental plant researchers". In *E-Science-Tage 2021: Share Your Research Data,* edited by Vincent Heuveline and Nina Bisheh, 132–145. heiBOOKS. DOI: https://doi.org/10.11588/heibooks.979.c13724.

NFDI4Plants. 2023a. "Annotated Research Contexts specification". Visited on May 30, 2023. https://github.com/nfdi4plants/ARC-specification.

———. 2023b. "ARC Ontology". Visited on May 30, 2023. https://github.com/nfdi4plants/ARC_ontology.

———. 2023c. "NFDI4Plants Ontology - An intermediate ontology for plants used by DataPLANT to fill the ontology gap". Visited on May 30, 2023. https://github.com/nfdi4plants/nfdi4plants_ontology.

———. 2023d. "Swate (Excel Add-In for annotation of experimental data and computational workflows". Visited on May 30, 2023. https://github.com/nfdi4plants/Swate.

———. 2023e. "Swate Templates". Visited on May 30, 2023. https://github.com/nfdi4plants/Swate-templates.

———. 2023f. "Swobup". Visited on May 30, 2023. https://github.com/nfdi4plants/Swobup.

Papoutsoglou, Evangelia A., Daniel Faria, Daniel Arend, Elizabeth Arnaud, Ioannis N. Athanasiadis, Inês Chaves, Frederik Coppens, et al. 2020. "Enabling reusability of plant phenomic datasets with MIAPPE 1.1". *New Phytologist* 227 (1): 260–273. DOI: https://doi.org/10.1111/nph.16544.

Verein Nationale Forschungsdateninfrastruktur (NFDI) e.V. 2023. "Nationale Forschungsdaten Infrastruktur (NFDI)". Visited on September 4, 2023. https://www.nfdi.de/.

Walls, Ramona L., Balaji Athreya, Laurel Cooper, Justin Elser, Maria A. Gandolfo, Pankaj Jaiswal, Christopher J. Mungall, et al. 2012. "Ontologies as integrative tools for plant science". *American Journal of Botany* 99 (8): 1263–1275. DOI: https://doi.org/10.3732%2Fajb.1200222.

Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersbergand, Gabrielle Appleton, Myles Axtonand, Arie Baakand, Niklas Blombergand, et al. 2016. "The FAIR Guiding Principles for scientific data management and stewardship". *Scientific data* 3 (1): 1–9. DOI: https://doi.org/10.1038/sdata.2016.18.