Improving Metadata Collection and Aggregation in Plant Phenotyping Experiments with MIAPPE Wizard and DataPLANT

Daniel Arend[1], **Sebastian Beier**[2], **Dominik Brilhaus**[3], **Hannah Dörpholz**[2], **Manuel Feser**[1], **Kevin Frey**[4], **Patrick König**[1], **Oliver Maus**[1], **Dennis Psaroudakis**[1], **Cristina Martins Rodrigues**[5], **Andrea Schrader**[6], **Elisa Senger**[2], **and Heinrich Lukas Weil**[4]

1 Leibniz Institute for Plant Genetics and Crop Plant Research (IPK) Gatersleben, Germany **2** Forschungszentrum Juelich, CEPLAS, BioSC, Institute of Bio- and Geosciences, IBG4 Bioniformatics, 52428 Juelich, Germany **3** Cluster of Excellence on Plant Sciences (CEPLAS) / Heinrich-Heine-University Düsseldorf, Germany **4** RPTU Kaiserslautern-Landau, Germany **5** Albert-Ludwigs-Universität Freiburg, Germany **6** Cluster of Excellence on Plant Sciences (CEPLAS) / University of Cologne, Germany

BioHackathon series:
BioHackathon Germany 2022
Lutherstadt Wittenberg, Germany, 2022
Project 1/8

Submitted: 05 Apr 2023

Introduction

As part of the BioHackathon Germany 2022, we hereby report on the success of the two projects “MIAPPE Wizard: Enabling easy creation of MIAPPE-compliant ISA metadata for Plant Phenotyping Experiments” and “DataPLANT - Facilitating Research Data Management to combat the reproducibility crisis”. Shortly before the actual hackathon, it became apparent to the participants that close coordination between the projects would be very beneficial. Both projects aimed to improve the process of collecting and aggregating metadata on plant experiments, but with different approaches. In the following, we summarize the accomplished work and discuss future developments.

Back in 2015 (Krajewski et al., 2015), the plant community came together to discuss recommendations on how to best capture plant phenotyping experiments in regards to the data and associated metadata. After another year of discussion and looking at different use cases, the formal “minimal information about plant phenotyping experiment” (short MIAPPE, see also https://miappe.org/) with a reference implementation using ISA-Tab was published (Ćwiek-Kupczyńska et al., 2016). ISA-Tab is the file representation of the generic ISA (Investigation - Study - Assay) framework (Rocca-Serra et al., 2010), where experiments are hierarchically mapped into the three sub levels. In addition to the ISA-Tab representation of the ISA model, there is now also the ISA-JSON representation. Since 2020, the current version v1.1 of MIAPPE was released (Papoutsoglou et al., 2020). With that, its scope was extended to also include woody plants (version 1 primarily covered field crops) and was aligned to other metadata standards (DublinCore, Multi-Crop Passport Descriptors), APIs (BreedingAPI, GnpIS-Ephesis), ontologies (Crop Ontology) and include controlled vocabulary and ISO norms where possible. In their current form, the MIAPPE specifications cover a set of more than 90 core variables and a further, non-exhaustive list of 40 additional environmental variables and experimental factors.

Although many researchers in the plant community welcome and frequently cite this metadata standard, the number of publicly available MIAPPE-compliant datasets is very limited. Through conversations with biologists and data stewards, criticism is often voiced that there are far too many metadata fields demanding an intense investment of time and work to fill them. The reason often given is the time-consuming manual work that could be used with perceived more important activities, such as analysing the data.

MIAPPE Wizard - Enabling easy creation of MIAPPE-compliant ISA metadata for Plant Phenotyping Experiments

Experimental research data is only useful to the research community, if it is FAIR (Wilkinson et al., 2016). In the plant phenomics domain the MIAPPE standard was developed to provide a minimal checklist for relevant metadata attributes and a suitable data model to describe experiments in a FAIR way. To publish metadata following these guidelines, frameworks like ISA can be used. Nevertheless, the hurdle for the data producer to record and manage their experimental metadata is quite high, due to missing data management experience, a diverse set of tools and rare data stewardship support.

To overcome these challenges we decided to implement an intuitive graphical user interface, which guides data producers throught the process of describing their experiments. A step-by-step process supported by smart content recommendations and an appropriate formatting in parallel guarantee a sustainable metadata package without the need to be familiar with ISA/MIAPPE standards. This software was named MIAPPE Wizard and a first prototype has been deployed during the Hackathon.

DataPLANT - Facilitating Research Data Management to combat the reproducibility crisis

In pursuit of FAIR research data, the DataPLANT consortium has introduced the Annotated Research Context (ARC), a FAIR research object built upon existing standards like ISA, CWL and RO-Crate.

Significant effort has been invested in developing tools and documentation to support the creation of ARCs. During the BioHackathon Germany 2022, our focus was on further integrating these tools with the biological research community by providing our interfaces through openly available REST APIs. Additionally, we aimed to make the annotation and storing of experimental data as seamless as possible for researchers. To this end, we also enhanced the findability and quality of our training materials.

ARC Data Container

The ARC provides a comprehensive framework for describing the entire research cycle, from growing the plant to acquiring, processing, and publishing data. To achieve this, it builds upon existing standards such as ISA (Rocca-Serra et al., 2010), CWL (Amstutz et al., 2016), git (Blischak et al., 2016) and RO-Crate (Soiland-Reyes et al., 2022). ISA facilitates multi-faceted annotation of experimental workflows through its human- and machine-readable structure that can accommodate any information. The ARC extends this structure with git and CWL to support a diverse range of experimental workflows, including plant phenotyping.

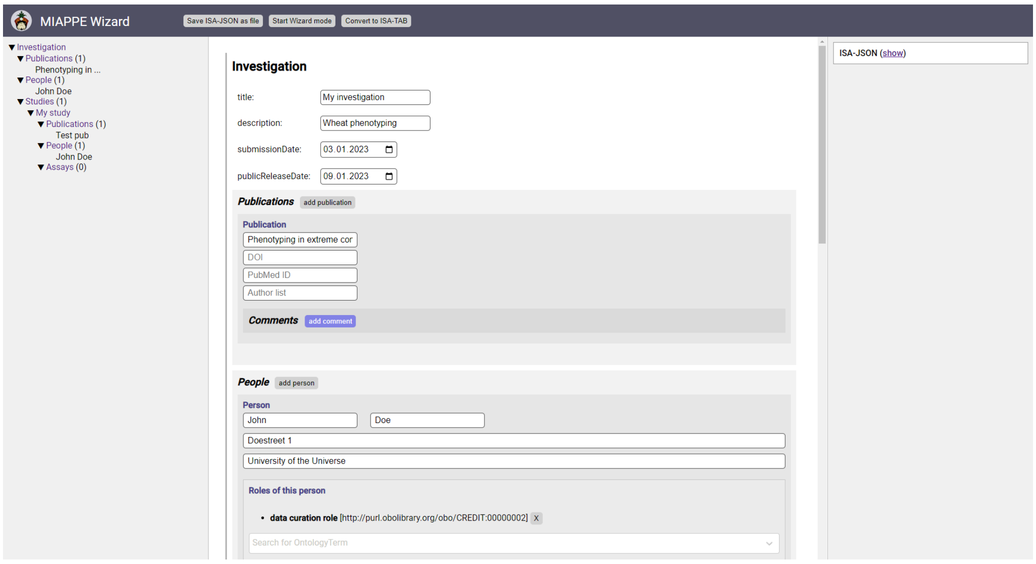
However, selecting relevant information to annotate a given experiment can be challenging. To address this challenge, checklists have been developed to help researchers create MIAPPE-compliant ARCs that capture and organize all necessary metadata in a structured manner, facilitating data discovery and reuse. Furthermore, the extensive nature of the MIAPPE standard provides a useful test case for evaluating the flexibility and scalability of the ARC data container as it must accommodate a large number of metadata fields and various types of experimental data.

Results

MIAPPE Wizard Prototype

A prototype of the MIAPPE Wizard has been developed using the Svelte framework (https://svelte.dev) as a single-page application. Upon launching the application, users can choose to start in either questionnaire or form mode or load a previously saved ISA-JSON file to continue editing. The interface is divided into three columns: the left column displays a tree navigation of the current ISA data structure (see Figure 1), the middle column presents forms for data entry, and the right column provides helpful explanations of the ISA or MIAPPE entities and attributes.

The ISA data model's JSON schemas serve as a single source of truth for constructing the ISA-JSON object structure from specific ISA entities. Multiple recurring form elements have been implemented generically and are automatically displayed in the appropriate order. The questionnaire mode supports multiple-choice questions and step-by-step data entry. Multiple-choice questions can be used to gather relevant information about the experimental setup and process before actual data entry. This information can be used to pre-fill certain data fields or simplify the data entry process by pre-filling form fields for ontology terms or hiding irrelevant sections.



title: My investigation

description: Wheat phenotyping

submissionDate: 03.01.2023

publicReleaseDate: 09.01.2023

Phenotyping in extreme cor

John Doe

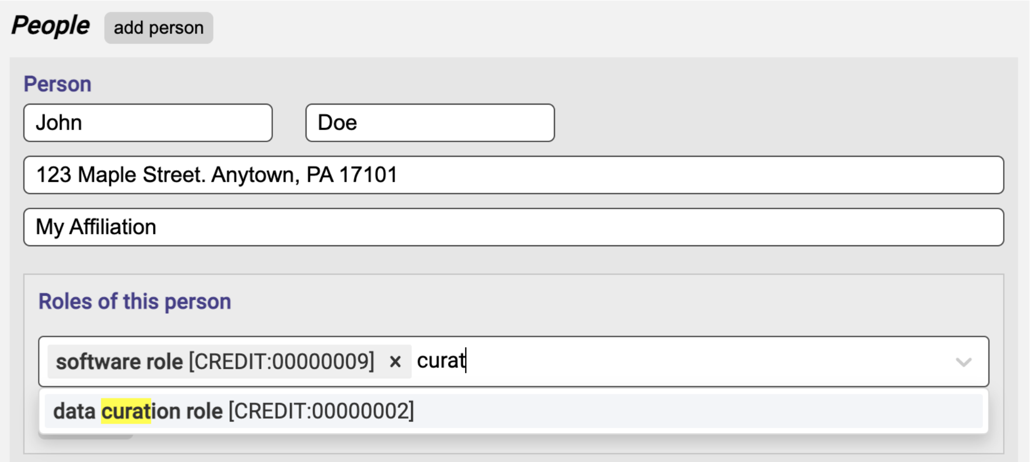
Doestreet 1

University of the Universe

- **data curation role** [http://purl.obolibrary.org/obo/CREDIT:00000002] X

Figure 1: Screenshot of the main layout of the MIAPPE Wizard

To enrich user input with ontology terms, the MIAPPE Wizard leverages DataPLANT's ontology lookup service through asynchrounous JavaScript requests in the background. This feature is implemented in a GUI component that allows users to interactively search a given ontology for matching terms, select and add desired terms from the suggestions (see Figure 2).



John

Doe

123 Maple Street. Anytown, PA 17101

My Affiliation

software role [CREDIT:00000009] × curat

data curation role [CREDIT:00000002]

Figure 2: GUI component to interactively look up and add ontology terms to the user's ISA objects.

Furthermore, we implemented a simple ISA-JSON to ISA-Tab conversion backend based on the python isatools-api, available at https://webapps.ipk-gatersleben.de/isa-json2tab/json2tab (source code at https://github.com/IPK-BIT/isa-json2tab). The MIAPPE Wizard can submit user generated ISA-JSON to this web service and receive the same information in ISA-Tab representation in response, which it then offers for download through the browser. This allows the Wizard to output metadata in both ISA-JSON and ISA-Tab formats.

As ISA-JSON and the form generated in the GUI can become large and overwhelming for users, it is essential to provide navigation assistance. In the MIAPPE Wizard this is achieved through a tree view that is familiar to users from many applications. The tree view allows subobjects with complex nesting, such as individual studies. For lists, such as persons, publications, and studies, the content is displayed but not clickable, allowing users to check for completeness at a glance. The length of the list is displayed next to the node. Clicking a node sets the focus of the Wizard to the selected element, hiding everything else. To return to the default view, users can click the Investigation node. To accommodate large ISA-JSON objects, the tree view is collapsible.

We decided to initially adapt a reduced set of the MIAPPE data model for the prototype of the MIAPPE Wizard that includes components such as Persons, Publications and Material. These components can be dynamically complemented in the future. We have discussed various solutions for data entry and implemented data type-specific GUI components. In the future, it is planned to continuosly improve the components and add further implementations to support the entire MIAPPE data model.

ArcCommander REST API

As part of our work on the DataPLANT project at the BioHackathon Germany 2022, we have made significant progress in consuming MIAPPE ISA-JSON generated by the MIAPPE Wizard into ARCs. For this we implemented an ArcCommander REST API and the ISA-JSON import functionality.

The initial development of the ArcCommander REST API has been completed and successfully tested in a local environment. The API allows users to interact with the ARC data container using standard RESTful API calls, facilitating integration with existing workflows and data management systems. The ArcCommander can be run in server mode, making its CLI-designed functionality accessible via REST API calls.

In addition to the ArcCommander REST API, we have also made significant progress on the ISA-JSON import backend. This component enables users to import existing experimental metadata into an ARC data container, streamlining the process of migrating data from existing data management systems into the ARC format. We have completed the initial implementation of the backend and have tested it with MIAPPE ISA-JSON test data generated using the official ISA-API.

DataPLANT Ontology Service

The DataPLANT ontology service comprises three main tool components. 1. Swate (https://github.com/nfdi4plants/Swate), an MS Excel add-in that provides a graphical user interface for incrementally creating ISA-tab like annotation tables with support for minimal information standards and an ontology term look up service. 2. SwateDB, a neo4j graph database containing minimal information standards and a preselection of useful ontologies. 3. Swobup (https://github.com/nfdi4plants/Swobup), which forwards community input to SwateDB, allowing live updates from multiple GitHub repositories directly into the database.

To support the prototyping of the MIAPPE Wizard, we improved the accessibility of the Swate ontology API by relaxing the CORS (Cross-Origin Resource Sharing) policy to allow external browser-based access. We also made minor form changes to conform to OpenAPI standards and created interactive documentation using SwaggerUI (https://swate.nfdi4plants.org/docs/IOntologyAPIv2.html). Finally, we modified some endpoints to accurately perform the specific search queries of the MIAPPE Wizard.

In order to expand the ontology service to cover more terms relevant in plant phenotyping experiments, we have included two additional ontologies: the CRediT ontology (Contributer Roles Taxonomy, https://credit.niso.org/), which contains 14 typical contributer roles to annotating person information within an investigation, and the MIAPPE ontology, which contains all terms from the MIAPPE data model v1.1 (https://github.com/MIAPPE/MIAPPE/tree/master/MIAPPE_Checklist-Data-Model-v1.1).

During the BioHackathon, we addressed several challenges, including correcting the format of the official CRediT .obo file to provide a functional input file for the ontology service and creating an .obo file of the MIAPPE ontology that was previously only available in OWL format. Although ontology format converters could have been used, we opted to create the OBO file based on the data model to avoid the errors that can arise from converting from OWL to OBO, especially when the OWL file contains datatype properties. Both ontologies were successfully incorporated into the ontology service.

In order to provide an interface within the DataPLANT project that facilitates the annotation of plant phenotyping experiments, we created multiple MIAPPE-compliant template files using the MS Excel add-in Swate. These templates will assist users in documenting metadata and data in MS Excel format by providing a list of relevant metadata and data terms that should be included in their experiment descriptions to make them FAIR. The templates cover different sections of the MIAPPE data model checklist v1.1 and are currently being reviewed to reference the newly added MIAPPE ontology. We also noted that the MIAPPE terms could be extended to reach a broader audience and cover more use cases.

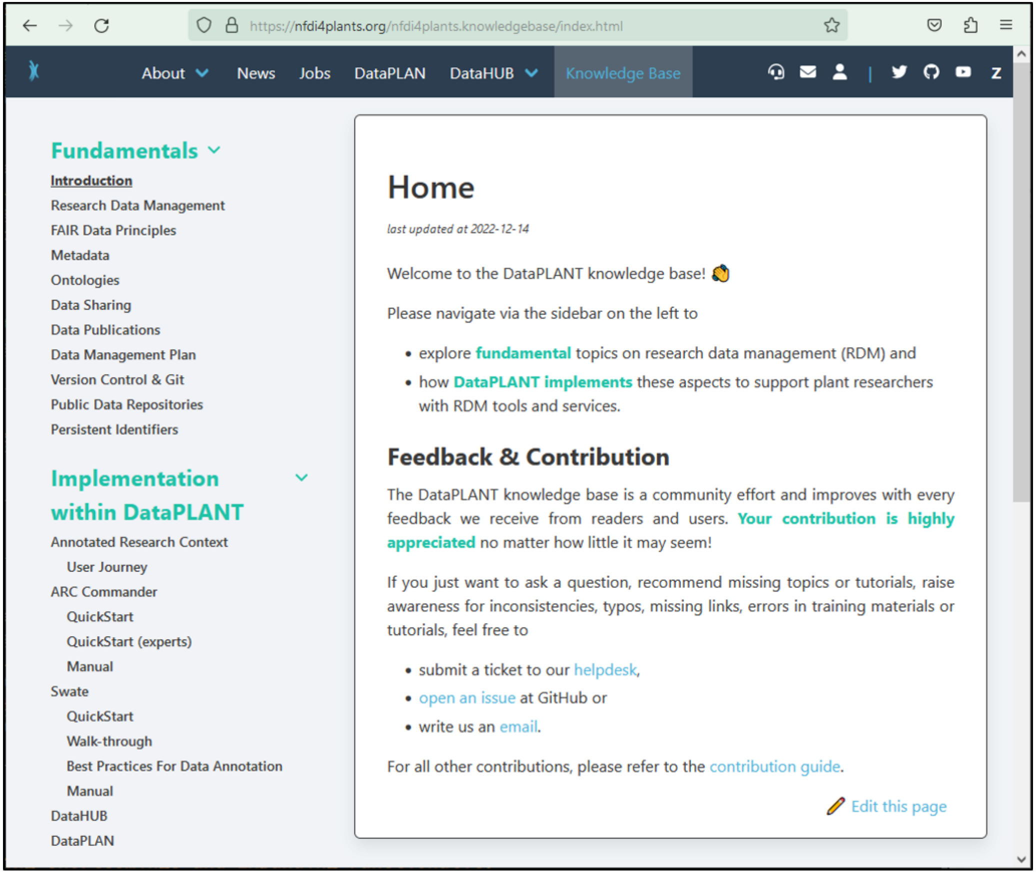
DataPLANT Tool Documentation

During the BioHackathon Germany 2022, we focused on consolidating scattered documentation, including a general description of the ArcCommander, wiki, quickstart for users, and developers' documentation from different GitHub repositories.

Now, a centralized ArcCommander collection is available in the DataPLANT knowledge base, providing a one-stop experience for gathering the information they need to rapidly create their first ARC, learn about the background of an ARC in detail, and access the ArcCommander developers' documentation.

This allows users to focus on working in the ARC and providing ontology-driven metadata annotation to address the reproducibility crisis(Baker, 2016), which was central to both projects presented here. An expert quickstart was added, proving to be a significant shortcut even for beginners who have created at least one ARC previously.

To balance speed and user-friendliness with structured information, we converted a wiki and additional information into a manual for the ArcCommander and embedded it into the DataPLANT Knowledge Base. We also overhauled the sidebar of the webpage to unify structure and names (e.g. quickstarts) and create a simpler visual appearance (see Figure 3).



Home

last updated at 2022-12-14

Welcome to the DataPLANT knowledge base! 👋

Please navigate via the sidebar on the left to

- explore **fundamental** topics on research data management (RDM) and
- how **DataPLANT implements** these aspects to support plant researchers with RDM tools and services.

Feedback & Contribution

The DataPLANT knowledge base is a community effort and improves with every feedback we receive from readers and users. **Your contribution is highly appreciated** no matter how little it may seem!

If you just want to ask a question, recommend missing topics or tutorials, raise awareness for inconsistencies, typos, missing links, errors in training materials or tutorials, feel free to

- submit a ticket to our helpdesk,
- open an issue at GitHub or
- write us an email.

For all other contributions, please refer to the contribution guide.

✏ Edit this page

Figure 3: Screenshot of the DataPLANT Knowledge Base landing page with the sidebar and link to the contribution guide and tool-centred centralized reorganization (2023-02-17)

To address frequently asked questions from the community and hands-on sessions, we initiated a new FAQ section that is currently under development, with questions provided by attendees.

From the outset, the community was invited to contribute to the DataPLANT Knowledge Base through an "edit this page" option located at the bottom of each page. This link leads to the document in the GitHub directory, where it can be edited and a pull request created.

In order to consolidate contribution routes and to encourage community members at all levels to contribute to the Knowlege Base, we launched a more detailed contribution guide during the Hackathon. Multiple links were adjusted, explanations and examples were expanded (e.g., how to construct an appropriate relative path), and a link was provided explaining how to test changes or new contributions in a local environment before submitting a pull request. The contribution guide can currently be accessed via a link on the Knowledge Base landing page.

The kick-off for the online version of this contribution guide complements the collaborative community approach of the MIAPPE and DataPLANT projects at the BioHackathon Germany 2022.

Future Tasks: Addressing Shortcomings and Expanding Functionality

During discussions with attendees and virtual participants, it became clear that the ISA representation of the MIAPPE data model still has some shortcomings. We have made suggestions to address these issues and plan to improve the mapping in the coming months. The MIAPPE Wizard prototype has already demonstrated some comprehensive and useful features. In the next few months, we will work to integrate the complete ISA model and full MIAPPE checklist to enable the Wizard to produce reusable metadata.

The MIAPPE Wizard project has already leveraged valuable DataPLANT services, such as the ontology lookup, to enhance the functionality of its user interface. In the future, we plan to expand this connection and utilize additional services, such as the ArcCommander for exporting ARC containers in addition to ZIP files.

Acknowledgements

This work was funded by ELIXIR, the research infrastructure for life-science data, by the Federal Government of Germany and the county of North Rhine-Westphalia (de.NBI - the german Network for Bioinformatics Infrastructure), and the DFG in frame of the NFDI4BioDiversity and DataPLANT consortia of the NFDI. Discussions and work were conducted during the 1. BioHackathon Germany in Lutherstadt Wittenberg, Germany. We also thank the remote participants Timo Mühlhaus, Cyril Pommier, Marko Petek, Maja Zagorščak, and Philippe Rocca-Sera for a lively and fruitful discussion.

References

Amstutz, P., Crusoe, M. R., Nebojša Tijanić, Chapman, B., Chilton, J., Heuer, M., Kartashov, A., Leehr, D., Ménager, H., Nedeljkovich, M., Scales, M., Soiland-Reyes, S., & Stojanovic, L. (2016). *Common workflow language, v1.0*. figshare. https://doi.org/10.6084/M9.FIGSHARE.3115156.V2 **[cito:citesAsAuthority]**

Baker, M. (2016). 1, 500 scientists lift the lid on reproducibility. *Nature*, *533*(7604), 452–454. https://doi.org/10.1038/533452a **[cito:citesAsAuthority]**

Blischak, J. D., Davenport, E. R., & Wilson, G. (2016). A quick introduction to version control with git and GitHub. *PLOS Computational Biology*, *12*(1), e1004668. https://doi.org/10.1371/journal.pcbi.1004668 **[cito:citesAsAuthority]**

Ćwiek-Kupczyńska, H., Altmann, T., Arend, D., Arnaud, E., Chen, D., Cornut, G., Fiorani, F., Frohmberg, W., Junker, A., Klukas, C., Lange, M., Mazurek, C., Nafissi, A., Neveu, P., Oeveren, J. van, Pommier, C., Poorter, H., Rocca-Serra, P., Sansone, S.-A., … Krajewski, P. (2016). Measures for interoperability of phenotypic data: Minimum information requirements and formatting. *Plant Methods*, *12*(1). https://doi.org/10.1186/s13007-016-0144-4 **[cito:citesAsAuthority]**

Krajewski, P., Chen, D., Ćwiek, H., Dijk, A. D. J. van, Fiorani, F., Kersey, P., Klukas, C., Lange, M., Markiewicz, A., Nap, J. P., Oeveren, J. van, Pommier, C., Scholz, U., Schriek, M. van, Usadel, B., & Weise, S. (2015). Towards recommendations for metadata and data handling in plant phenotyping. *Journal of Experimental Botany*, *66*(18), 5417–5427. https://doi.org/10.1093/jxb/erv271 **[cito:citesAsAuthority]**

Papoutsoglou, E. A., Faria, D., Arend, D., Arnaud, E., Athanasiadis, I. N., Chaves, I., Coppens, F., Cornut, G., Costa, B. V., Ćwiek-Kupczyńska, H., Droesbeke, B., Finkers, R., Gruden, K., Junker, A., King, G. J., Krajewski, P., Lange, M., Laporte, M.-A., Michotey, C., … Pommier, C. (2020). Enabling reusability of plant phenomic datasets with MIAPPE 1.1. *New Phytologist*, *227*(1), 260–273. https://doi.org/10.1111/nph.16544 **[cito:citesAsAuthority]**

Rocca-Serra, P., Brandizi, M., Maguire, E., Sklyar, N., Taylor, C., Begley, K., Field, D., Harris, S., Hide, W., Hofmann, O., Neumann, S., Sterk, P., Tong, W., & Sansone, S.-A. (2010). ISA software suite: Supporting standards-compliant experimental annotation and enabling curation at the community level. *Bioinformatics*, *26*(18), 2354–2356. https://doi.org/10.1093/bioinformatics/btq415 **[cito:citesAsAuthority]**

Soiland-Reyes, S., Sefton, P., Crosas, M., Castro, L. J., Coppens, F., Fernández, J. M., Garijo, D., Grüning, B., Rosa, M. L., Leo, S., Carragáin, E. Ó., Portier, M., Trisovic, A., Community, R.-C., Groth, P., & Goble, C. (2022). Packaging research artefacts with RO-crate. *Data Science*, *5*(2), 97–138. https://doi.org/10.3233/ds-210053 **[cito:citesAsAuthority]**

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., Silva Santos, L. B. da, Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. (2016). The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, *3*(1). https://doi.org/10.1038/sdata.2016.18 **[cito:citesAsAuthority]**