The concept of SciWIn as part of the reproducible science toolset in FAIRagro

Harald von Waldow [1], Jens Krumsieck [1], Antonia Leidel [2] and Patrick König [2]

[1] Johann Heinrich von Thünen Institute, Braunschweig
[2] Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), Gatersleben

Thematic note

This text is deliverable D4.4.1 of Measure 4.4. in Task Area 4 of the NFDI consortium FAIRagro. "SciWIn" stands for **Sci**entific **W**orkflow **In**frastructure and denotes the overall deliverable of Measure 4.4. This document concludes Action 1 of Measure 4.4. The title set forth in the proposal (Ewert et al., 2023) was "Joint concept of SciWIn as part of the RDC semantic toolset" Several assumptions made at the time of writing the proposal did not materialize. It was therefore necessary to adapt the direction of the project and consequently the thrust of its conceptualization.

The missing Research Data Commons

The proposal foresaw the integration of SciWIn into a joint infrastructure involving in particular an "RDC mediation layer" (Ewert et al., 2023), where "RDC" stands for "Research Data Commons". RDC was anticipated to become "an overarching virtual expandable infrastructure" (Glöckner et al., 2020) hosting "cross-cutting services for the NFDI" (Bierwirth et al., 2020). While Glöckner et al. (2020) and Bierwirth et al. (2020) are mere declarations of intent, the consortium NFDI4BioDiversity proposed to establish RDC as a cloud-based research infrastructure and provided a high-level architectural layered concept for RDC (Glöckner et al., 2019) into which SciWIn was supposed to be integrated.

In addition to "RDC" as infrastructure, the term "RDC" was also used in the FAIRagro proposal in the sense of a set of criteria that services should fulfill be be inter-operable with the NFDI-wide infrastructure. It was proposed that "FAIRagro will comply wit h the NFDI-RDC" and that "Storage Instances [of M4.4] . . . will hold RDC-compliant FAIR DOs . . . "

In June 2024 we organized a meeting with stakeholders from NFDI4Biodiversity, which w ere involved in the design and implementation of the NFDI4BioDiversity-specific RDC (Bio-RDC). It turned out that at that point in time, RDC existed as "a blueprint", an "architectural model" and a collection of specific individual services, namely

- An image annotation software, BIIGLE
- The terminology-related BiodivPortal (not reachable at the time of writing)

- A search engine for biological data GFBio Search
- An object storage technology, Aruna
- An AAI-provider, Life Science Login by EOSC-Life
- A KPI monitoring service, Scorpion

The conceptual ideas had no actionable specification or reference implementation and also seemed to be still in flux. A move to a more domain-oriented decentralized architectural paradigm ("data mesh concept") was considered. The six approved R DC services on the other hand did not bear direct touching points with SciWIn or FAIRagro. A list of criteria that services should fulfill in order to be "RDC compliant" was planned by TA4 of NFDI4Biodiversity but not yet published.

Changed Directions

Under these circumstances we had to deal with the fact that there did not exist an "RDC" or an "RDC semantic toolset" into which SciWIn could have been meaningfully integrated. Also the realization of such a thing did not seem likely in a time-frame that would allow to take it into account in the planning and design of SciWIn. Another consequence of the lack of RDC or a clear path towards its realization was the lack of FAIRagro external cooperation partners to develop a "joint concept" with respect to "the RDC semantic toolset".

In order to stay true to the spirit of this action as originally considered, we developed a set of goals to capture the essence of Action 1. Drawing from Bierwirth et al. (2020), Glöckner et al. (2020) and Diepenbroek et al. (2023), we defined characteristics that would 1) increase the chance for SciWIn to become a part of an NFDI-RDC if that concept were actually implemented, and 2) maximize cross-domain usage, usefulness and synergies in any case:

1. While primarily use-cases and requirement from the agrosystem science community drive SciWIn's development in FAIRagro, it should nevertheless be domain-agnostic and potentially unleash its full potential also in other quantitative research domains.
2. SciWIn tools and services should be easily accessible for researchers from other NFDI consortia and other domains in general.
3. SciWIn should be based on data formats and protocols that are well established, domain-agnostic and future-proof to maximize the chances for wide adoption and interoperability.
4. SciWIn should actively exploit existing services and services that are currently being developed

    - to avoid re-invention of the wheel,
    - to increase development efficiency and
    - to be exposed early on to the greater research ecosystem

SciWIn components - Overview

The original idea of the SciWIn design, as laid out in an ecosystem map (Ewert et al., 2023, Figure 16) features five components:

1. An AAI provider,
2. The "Workflow Hub",
3. Compute instances,

4. Storage instances, and
5. Workflow Objects

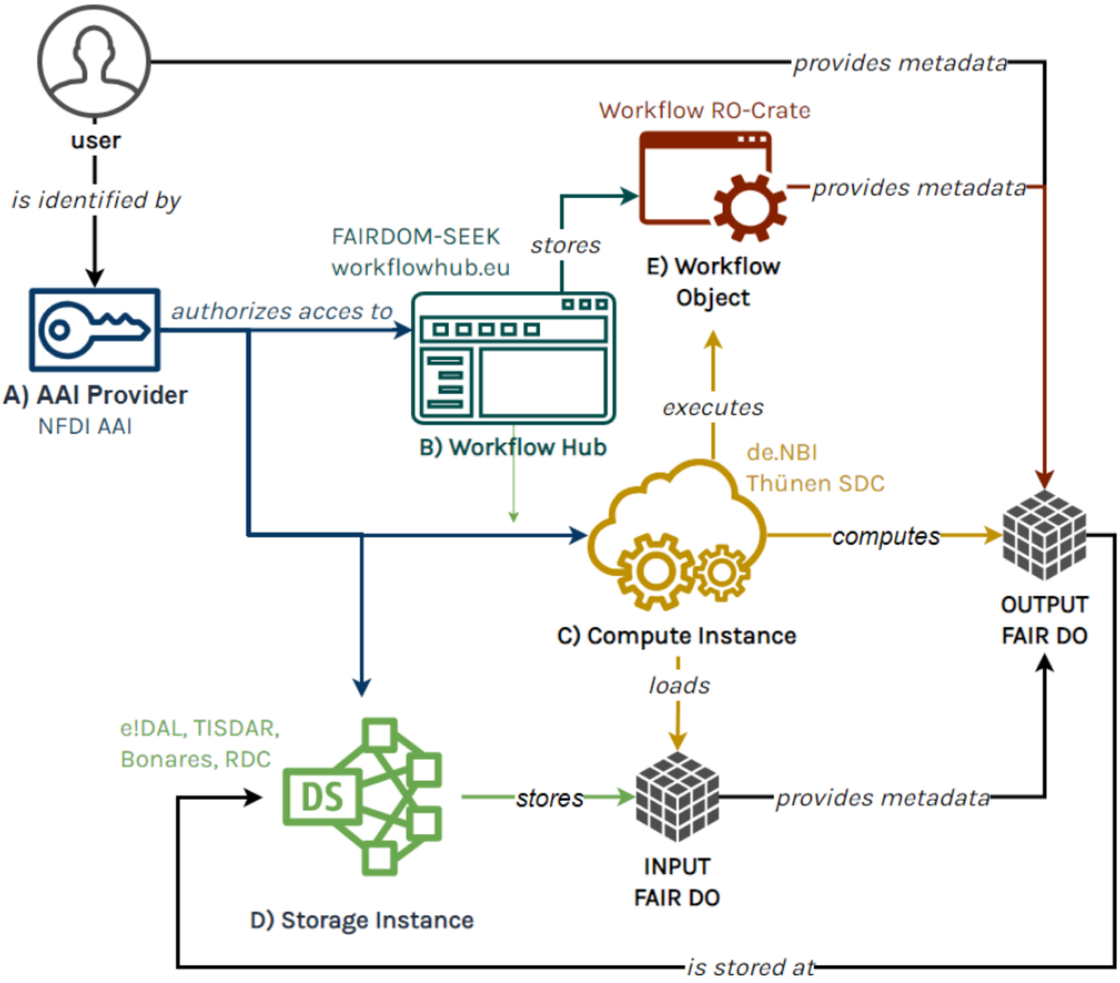

Figure 1: Original ecosystem idea for SciWIn (Figure 16 from Ewert et al., 2023)

Nomenclature refinements

In that conceptualization, only the “Workflow Hub” was supposed to be developed as a dedicated infrastructure item by SciWIn, while the other components are existing services that communicate with the “Workflow Hub”. The main purpose of the “Workflow Hub” was the provision of “an easy-to-use interface to work on and create new FAIR DO outputs with automatically annotated provenance graphs”. “FAIR DO” stands for “FAIR Digital Objects”, which is used for a quite abstract concept in the current literature. Schultes & Wittenburg (2019) state that FAIR Digital Objects “represent data, software or other research resources” and “must be accompanied by persistent identifiers, metadata and contextual documentation to enable discovery, citation and reuse”. We slightly modify and sharpen and the meaning of the terms “FAIR DO” and “Workflow Object” for use in SciWIn and at the same time adapt their semantics to better fit the current implementation strategy:

FAIR DO: A serializable object that adheres to the definition of Schultes & Wittenburg (2019) above. This also implies that a FAIR Do resides in a suitable FAIR repository that provides discovery and citeability.

Workflow Object: A data structure that holds a definition of a computational workflow, associated data and software or pointers to them, along with provenance and version information for all these objects. The Workflow Object can be consumed by an execution engine, which then might return the Workflow Object amended with results of the execution of a computational workflow.

Re-conceptualization of the "Workflow Hub"

Realizing that the main challenge to be solved lies in the provisioning of tooling for the easy creation of workflows, this task is now assigned to a stand-alone program that scientists use at their workstations in their daily habitual work without requiring internet-access, a central service, or authorization. This stand-along program is called **SciWIn-Client**. The program supports not only the creation but also the management of all aspects of Workflow Objects containing computational workflows. In particular, it facilitates sharing of and collaboration on Workflow Objects by providing import and export functions to suitable platforms.

The second important function of SciWIn-Client is the communication with compute instances to enable scientists to submit computational workflows for remote execution and fetch the results. SciWIn-Client thus implements the functionality that was assigned to "Workflow Hub" in the initial sketch in the proposal.

As a program that anybody can install on their computer, SciWIn-Client does not need an authentication service.

Re-conceptualization of "Storage Instances"

The "Storage Instances" mentioned in the proposal that still exist are e!DAL-PGP and the BonaRes Repository. Those are established research repositories, just like OpenAgrar or Zenodo, that serve a wide range of research communities and users and have their own set of challenging requirements. "Storage Instances" in that sense are called "Repositories" from here on. They are run and operated by independent entities who in general have no interest to invest resources into fulfilling very specific requirements of FAIRagro. Therefore they are not suited to FAIRly realize the full potential of re-usable, re-combineable, modular computational workflows. The existing repositories are still useful in this context to publish workflows as citeable scientific output that is reliably preserved over long time-spans. However, a programmatically driven, non interactive submission of content is not possible with such repositories, and sometimes even reading data requires interactive operation. We therefore refrain from a tight technical integration of such repositories into SciWIn. We do expect that users search and find data and code in such repositories, ideally even packaged as a FAIR DO that can be consumed directly by SciWIn. Search of and access to some of those repositories, covering specific needs of FAIRagro and the agrosystem research community is provided by the products of M4.2 and M4.3, the "Middleware" and the "Search Service", respectively. We consider features for SciWIn-Client that ease the publication of Workflow Objects to such repositories by providing prompting for required metadata and specific formatting of such metadata for selected repositories. In that context we also consider more domain and/or workflow specific repositories such as Workflow Hub or ARChive.

While such repositories are suited to publish FAIR DOs which have reached a certain level of quality, are sufficiently annotated with metadata and should be re-usable by researchers outside the lab or project where they originated, our idea of SciWIn also includes a way to share workflows, or more precisely Workflow Objects in a more ad-hoc, intermittent, less formal, easy manner that supports seamless cooperation but doesn't aim at producing citeable scientific output. As will become clear in the following section, any webservice that provides access to Git repositories will work for that purpose. Institutional installations of services such as GitLab or Forgejo, or even commercial platforms such as Bitbucket, GitLab.com or GitHub can be used. We aim in particular to make DataPLANT's PLANTdataHUB (Weil et al., 2023) available for the SciWIn-based collaboration on computational workflows.

Concretization of "Workflow Objects"

Workflow objects in the SciWIn-context are data structures that encapsulate the definition of workflows with associated code and data or references to code and data. Since a close collaboration between FAIRagro and NFDI Consortium DataPLANT is established on different levels, we have taken into account their version of a Workflow Objects, the **Annotated Research Context (ARC)** (DataPLANT, 2025), and the established standards on which it is based. These are the **Common Workflow Language** (Crusoe et al., 2022) to specify computational workflows and the **Research Object Crate** (RO-Crate, Soiland-Reyes et al., 2022) as a data structure to package data (which here also includes code and workflow descriptions) into which ARCs can be converted. While compatibility with the advanced tooling and infrastructure of DataPLANT is an important piece to achieve synergies with this consortium covering a neighboring research domain, CWL and RO-Crate represent the state-of-the art for workflow descriptions and semantically annotated metadata formats. Therefore, they are also used or considered by other NFDI consortia, e.g. NFDI4Ing (Bronger et al., 2022) and NFDI4Health (Löbe & Turner, 2024). Furthermore, the semantic annotation of metadata allows for the integration of such Workflow Objects into knowledge graphs that interconnect different domains.

ARCs are Git repositories and therefore contain the version history of code, data and metadata. While ARCs allow for workflow representations as CWL, they do not require the representation of the full provenance information of a workflow's elements. M4.4 aims to work with DataPLANT towards a common ARC specification and its corresponding RO-Crate profile, so that both, DataPLANT's compatibility requirements with related standards such as the ISA-model (Sansone et al., 2016) and the usability and workflow-representation requirements of SciWIn are met.

Implementation of Compute Instances

The choice of CWL as workflow description language ensures that workflows created by SciWIn-Client can be executed on a broad range of platforms (CWL community, 2025). However, many of these platforms require significant resources for setup and operation. Different platforms have different sets of compute back-ends, such as HTCondor, AWS, Azure, SLURM and Kubernetes. Additionally, the interaction with remote **compute instances** differs from platform to platform. Therefore, in order to experiment with remote execution of workflows and be able to pilot the whole range of SciWIn functionality, we have settled on **Reana** (Šimko et al., 2019) as primary execution platform. Originating from CERN, Reana is widely used and under active development since 8 years. Our main reason to settle on Reana however was the fact that NFDI4PUNCH provided us simple access to an instance at the Leibniz-Institute for Astrophysics Potsdam through an informal collaboration. Furthermore, the BASE4NFDI project MC4NFDI (*A Multicloud Infrastructure for the NFDI*) would have ensured robust, well-integrated access to Reana clusters for users of SciWIn. Unfortunately, the MC4NFDI proposal was rejected in the 7^{th} submission round.

To create robust and reliable access to compute instances that can scale to real-world workloads and be used by FAIRagro-associated researchers, we deploy and configure our own Reana-Installation on a Kubernetes cluster in the de.NBI-cloud. We also set up and configure the Kubernetes cluster to fully control and be able to experiment with the remote execution feature of SciWIn-Client.

The role of Authentication & Authorization

Access to Reana and other services, such as the *FAIRagro Searchable Inventory of Services and Data* (Ewert et al., 2023, pp. 94–96) will be managed by the FAIRagro Community AAI based on Keycloak. Availability to the wider research community will be realized through the NFDI-wide Base4NFDI project IAM4NFDI, that is supported by the Working Group Identity and Access Management (Pempe & Politze, 2022). Integration of this AAI solution into FAIRagro is performed by FAIRagro Measure 4.2. **SciWIn-Client** will support the respective authorization protocol to allow for the seamless remote execution of computational workloads for authorized users.

Current requirements

In proposal driven software development, the textbook recipes for requirements engineering do not work out of the box, since the usual stakeholder-structure is not present. In particular, a priori there are no “users” or “customers” that could be queried for requirements. Consequently, the acquisition of users has to be part of the project.

The overarching goal or business case of SciWIn as specified by the FAIRagro-proposal (Ewert et al., 2023) is it to “promote FAIR RDM” by facilitating the reproducibility, the deployment and the publication of data analysis workflows and simulation models.

FAIRagro-intrinsic requirements

Sources for intrinsic requirements are: - Measure-internal brainstorming regarding personal goals of the developers - regular meetings with colleagues in Task Area 4 - meetings with colleagues from other Task Areas - feedback from the Community Advisory Board

These requirements are not elicited in a systematic fashion, are expressed subjectively by various individuals and might change over time or assume changing priorities. At the time of writing we identify the following non-functional requirements for the software-output of Measure 4.4:

1. demonstration of a PoC
2. attract real use cases
3. create synergies with other consortia and other external projects
4. generate publications
5. adoption of the software in multiple domains
6. demonstration of research output that was made possible by use of SciWIn

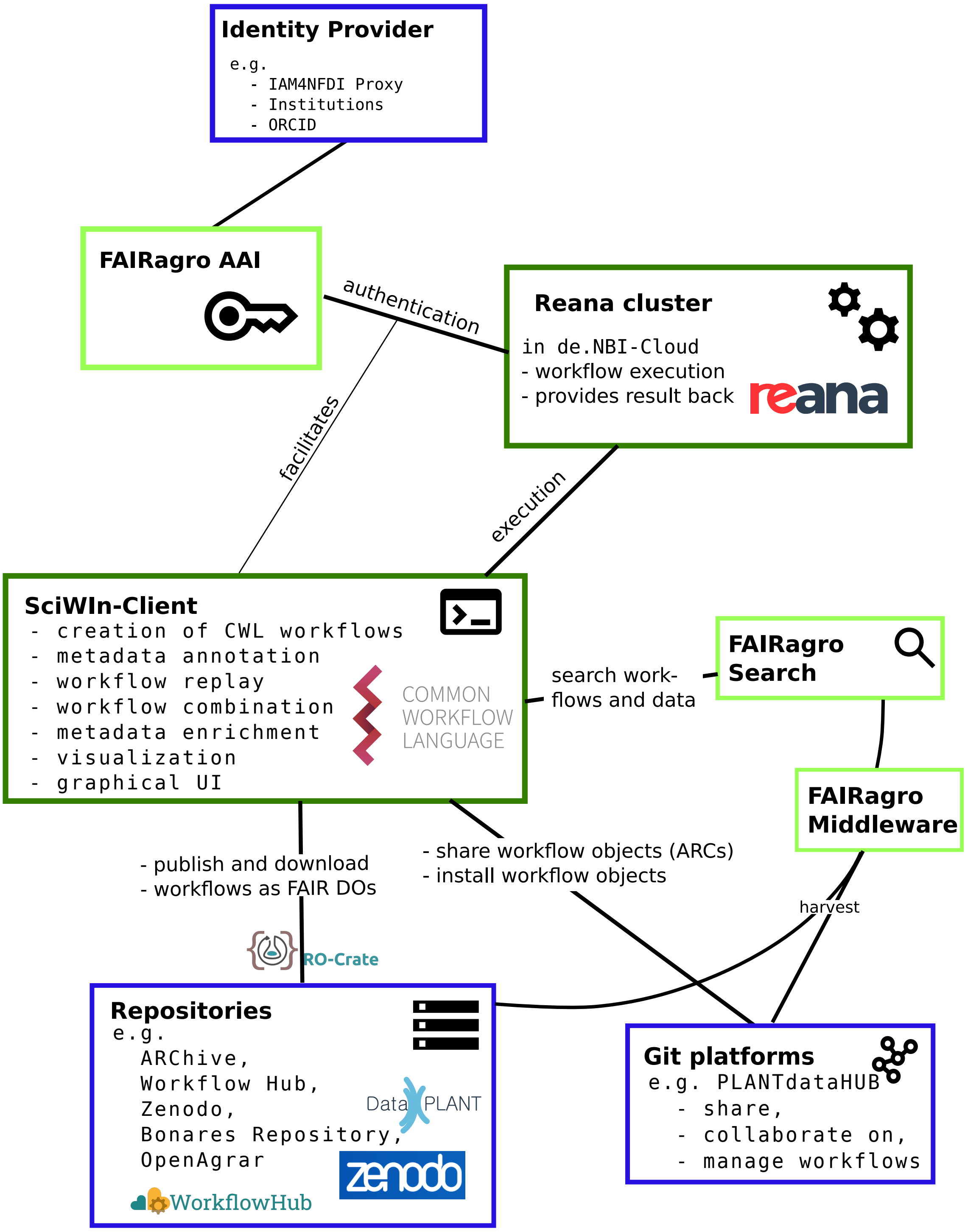

Figure 2: New ecosystem sketch

User requirements

In lieu of collecting requirements from the future users that can be found through a customer relationship in traditional software engineering (see e.g. Robertson et al., 2024), we initially relied on requirements that we devise ourselves, based on our personal and consulting experience with regard to scientific computing and research data management. Since then we could augment these requirements with information from users and potential users gathered at events such as the FAIRagro Plenary.

The currently considered requirement of SciWIn-Client can be roughly summarized as follows. Note however, that we follow an agile approach and continuously gather feedback from users and potential users. Prioritization of features therefore changes as as project progresses

- create machine readable workflow descriptions in a user-friendly manner
- integrate naturally into a common command-line interface oriented style of interactive work
- have a low threshold of learning before scientific work efficiency increases
- provide an easy overview of numerous versions, runs, inputs, outputs
- provide an easy way to annotate workflows for re-use
- provide an easy way for ad-hoc sharing of workflows
- allow to annotate workflows for publication
- provide a way to publish workflows as FAIR DOs in common repositories
- allow for local execution of workflows
- provide a frictionless way to access remote (powerful) computing resources
- record provenance information for workflows
- provide a graphical representation of workflows
- provide a graphical interface to manipulate workflows (connecting and splitting workflows)

Strategic requirements

The most salient non-functional requirements are:

- the need to fit into the FAIRagro landscape of tools and services, and
- the need to be compatible with the DataPLANT ecosystem of tools, services and data structures.

This section remains vague, because these requirements are in the process of being co-created with our colleagues in FAIRagro and DataPLANT.

State of the Implementation

SciWIn-Client has reached a state where it can be used independently by projects such as FAIRagro Use Cases, i.e. Use Case 7, NEXT-Gen-EXPERT. Version 1.0.0 was released Sep. 24, 2025. SciWIn-Client currently consists of over 15000 lines of Rust and has an extensive documentation and tutorial.

References

Bierwirth, M., Glöckner, F. O., Grimm, C., Schimmler, S., Boehm, F., Busse, C., Degkwitz, A., Koepler, O., & Neuroth, H. (2020, June 15). *Leipzig-Berlin-Erklärung zu NFDI-Querschnittsthemen der Infrastrukturentwicklung.* https://doi.org/10.5281/ZENODO.3895209

Bronger, T., Schlenz, H., Flemming, M., Selzer, M., & Jayavarapu, M. (2022). *SM4RO-c: SciMesh for RO-crate* (Version 1.0.0). Zenodo. https://doi.org/10.5281/zenodo.7414347

Crusoe, M. R., Abeln, S., Iosup, A., Amstutz, P., Chilton, J., Tijanić, N., Ménager, H., Soiland-Reyes, S., Gavrilović, B., Goble, C., & Community, T. C. (2022). Methods included: Standardizing computational reuse and portability with the Common Workflow Language. *Communications of the ACM*, *65*(6), 54–63. https://doi.org/10.1145/3486897

CWL community. (2025). *What can execute CWL descriptions?* https://www.commonwl.org/implementations

DataPLANT. (2025). *Nfdi4plants/ARC-specification: Annotated Research Context Specification v3.0-draft.2* (Version 3.0.0-draft.2) [Computer software]. Zenodo. https://doi.org/10.5281/ZENODO.8302661

Diepenbroek, M., Kostadinov, I., Seeger, B., Glöckner, F., Dieckmann, M., Goesmann, A., Ebert, B., Schimmler, S., & Sure-Vetter, Y. (2023). Towards a Research Data Commons in the German National Research Data Infrastructure NFDI: Vision, Governance, Architecture. *Proceedings of the Conference on Research Data Infrastructure*, *1*. https://doi.org/10.52825/cordi.v1i.355

Ewert, F., Specka, X., Anderson, J. M., Arend, D., Asseng, S., Boehm, F., Feike, T., Fluck, J., Gackstetter, D., Gonzales-Mellado, A., Hartmann, T., Haunert, J.-H., Hoedt, F., Hoffmann, C., König, P., Lesch, S., Lindstädt, B., Lischeid, G., Martini, D., . . . Weiland, C. (2023). *FAIRagro - A FAIR Data Infrastructure for Agrosystems (proposal).* https://doi.org/10.5281/ZENODO.8366884

Glöckner, F. O., Diepenbroek, M., Felden, J., Güntsch, A., Stoye, J., Overmann, J., Wimmers, K., Kostadinov, I., Yahyapour, R., Müller, W., Scholz, U., Triebel, D., Frenzel, M., Gemeinholzer, B., Goesmann, A., König-Ries, B., Bonn, A., & Seeger, B. (2019). *NFDI4BioDiversity - A Consortium for the National Research Data Infrastructure (NFDI).* https://doi.org/10.5281/ZENODO.3943645

Glöckner, F. O., Diepenbroek, M., Felden, J., Overmann, J., Bonn, A., Gemeinholzer, B., Güntsch, A., König-Ries, B., Seeger, B., Pollex-Krüger, A., Fluck, J., Pigeot, I., Toralf, K., Mühlhaus, T., Wolf, C., Heinrich, U., Steinbeck, C., Koepler, O., Stegle, O., . . . Bernard, L. (2020). *Berlin Declaration on NFDI Cross-Cutting Topics.* https://doi.org/10.5281/zenodo.3457213

Löbe, M., & Turner, D. (2024). *Applying FAIR signposting and RO-crate at NFDI4Health.* Zenodo. https://doi.org/10.5281/zenodo.14589428

Pempe, W., & Politze, M. (2022). *Concept for setting up a working group in the NFDI section "common infrastructures"* (Version 1.0). Zenodo. https://doi.org/10.5281/zenodo.6421866

Robertson, J., Robertson, S., & Reed, A. (2024). *Mastering the requirements process* (Fourth). Addison-Wesley.

Sansone, S.-A., Rocca-Serra, P., Gonzalez-Beltran, A., Johnson, D., & ISA Community. (2016). *Isa Model And Serialization Specifications 1.0.* https://doi.org/10.5281/ZENODO.163640

Schultes, E., & Wittenburg, P. (2019). FAIR Principles and Digital Objects: Accelerating Convergence on a Data Infrastructure. In *Data Analytics and Management in Data Intensive Domains: 20th International Conference, DAMDID/RCDL 2018, Russia, October 9–12, 2018, Reised Selected Papers* (Vol. 1003, pp. 3–16). Springer International Publishing. https://link.springer.com/10.1007/978-3-030-23584-0

Šimko, T., Heinrich, L., Hirvonsalo, H., Kousidis, D., & Rodríguez, D. (2019). REANA: A System for Reusable Research Data Analyses. *EPJ Web of Conferences*, *214*, 06034. https://doi.org/10.1051/

epjconf/201921406034

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

Weil, H. L., Schneider, K., Tschöpe, M., Bauer, J., Maus, O., Frey, K., Brilhaus, D., Martins Rodrigues, C., Doniparthi, G., Wetzels, F., Lukasczyk, J., Kranz, A., Grüning, B., Zimmer, D., Deßloch, S., von Suchodoletz, D., Usadel, B., Garth, C., & Mühlhaus, T. (2023). PLANTdataHUB: A collaborative platform for continuous FAIR data sharing in plant research. *The Plant Journal*, *116*(4), 974–988. https://doi.org/10.1111/tpj.16474