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 delivrable 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 fulfil be be interoperable with the NFDI-wide infrastructure. It was proposed that "FAIRagro will comply with 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 were 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 specfic 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 RDC services on the other hand did not bear direct touching points with SciWIn or FAIRagro. A list of criteria that services should fulfil 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 timeframe 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 developent 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 concortial 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 servcies 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

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

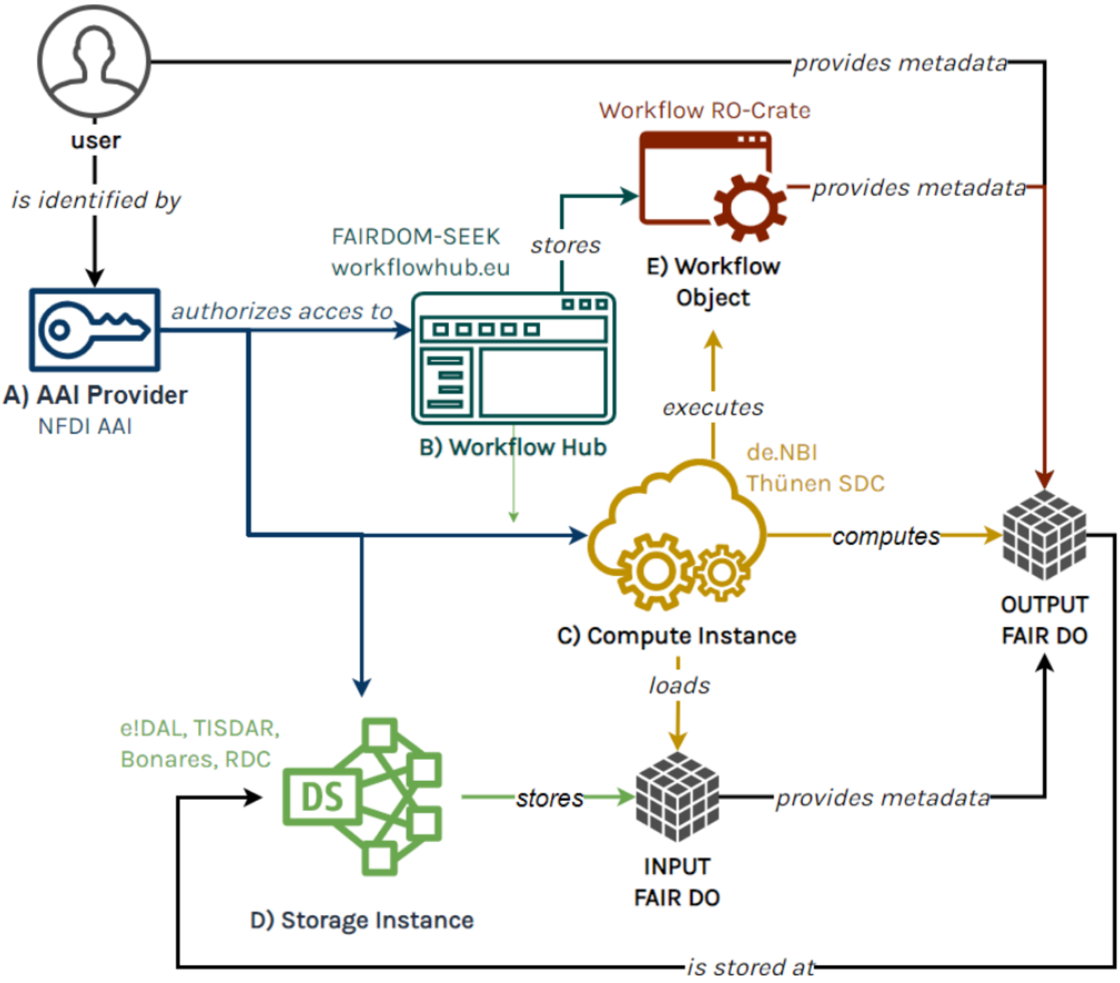

Figure 1: Original ecosystem idea for SciWIn (Figure 16 from Ewert et al., 2023)

In that conceptualization, only the "Workflow Hub" was supposed to be developed an 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 creation of "Workflow Objects". The actual desing at he time of writing differs from this early sketch. Figure 2 depicts the current high-level conceptualization of the SciWIn ecosystem.

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 workflow without requiring internet-access, a central service, or authorization. This stand-along program is called **SciWIn-Client**. 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.

In addition to SciWIn-Client we are planning to realize a second software-project within Measure 4.4, the **SciWIn-Hub**. The need for SciWIn-Hub stems from the realization that the data repositories ("Storage Instances" such as *e!DAL-PGP*, *Bonares*, *TISDAR*[1] in the proposal) are not suited to FAIRly to 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 programatically driven, non interactive submission of content is not possible with such repositories, and sometimes even reading data requires interactive operation.

Access to SciWIn-Hub and other services, such as the *FAIRagro Searchable Inventory of Services and Data* (Ewert et al., 2023, pp. 94–96) and compute instances, will be managed by 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 implement the respective authorization protocol.

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 FAIR Digital Object, the **Annotated Research Context (ARC)**, 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). 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 consortial, e.g. NFDI4Ing (Bronger et al., 2022) and NFDI4Health (Löbe & Turner, 2024). Furthermore, the semanitic annotation of metadata allows for the integration of such FAIR Digital Objects into knowledge graphs that interconnect different domains.

While we consider the adoption of CWL and the RO-Crate as serialization formats to produce FAIR Digital Objects (De Smedt et al., 2020) as future-proof and highly interopearable choice, true, actionable interoperability doesn't seem to be feasible with exactly one universal format. We therfore do not consider anymore a single specification of a "workflow object" as a deliverable of Measure 4.4. We rather expect to accept and produce a variety of formats to interact with the external infrastructures and services that turn out to be useful for SciWIn users.

The choice of CWL as workflow description language ensures that workflows created by SciWIn can be executed on a broad range of platforms (CWL community, 2025). However, many of these platforms require significant ressources for setup and operation. Different platforms have different sets of compute backends, 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 collaboaration. Furthermore, the BASE4NFDI project MC4NFDI (*A Multicloud Infrastructure for the NFDI*) would

[1] TISDAR is now called "Thünen-Atlas" and refers to the public repository of geospatial data at the Thünen Institute: atlas.thuenen.de.

have ensured robust, well-integrated access to Reana clusters for users of SciWIn. Unfortunately, the MC4NFDI proposal was rejected in the 7th submission round.

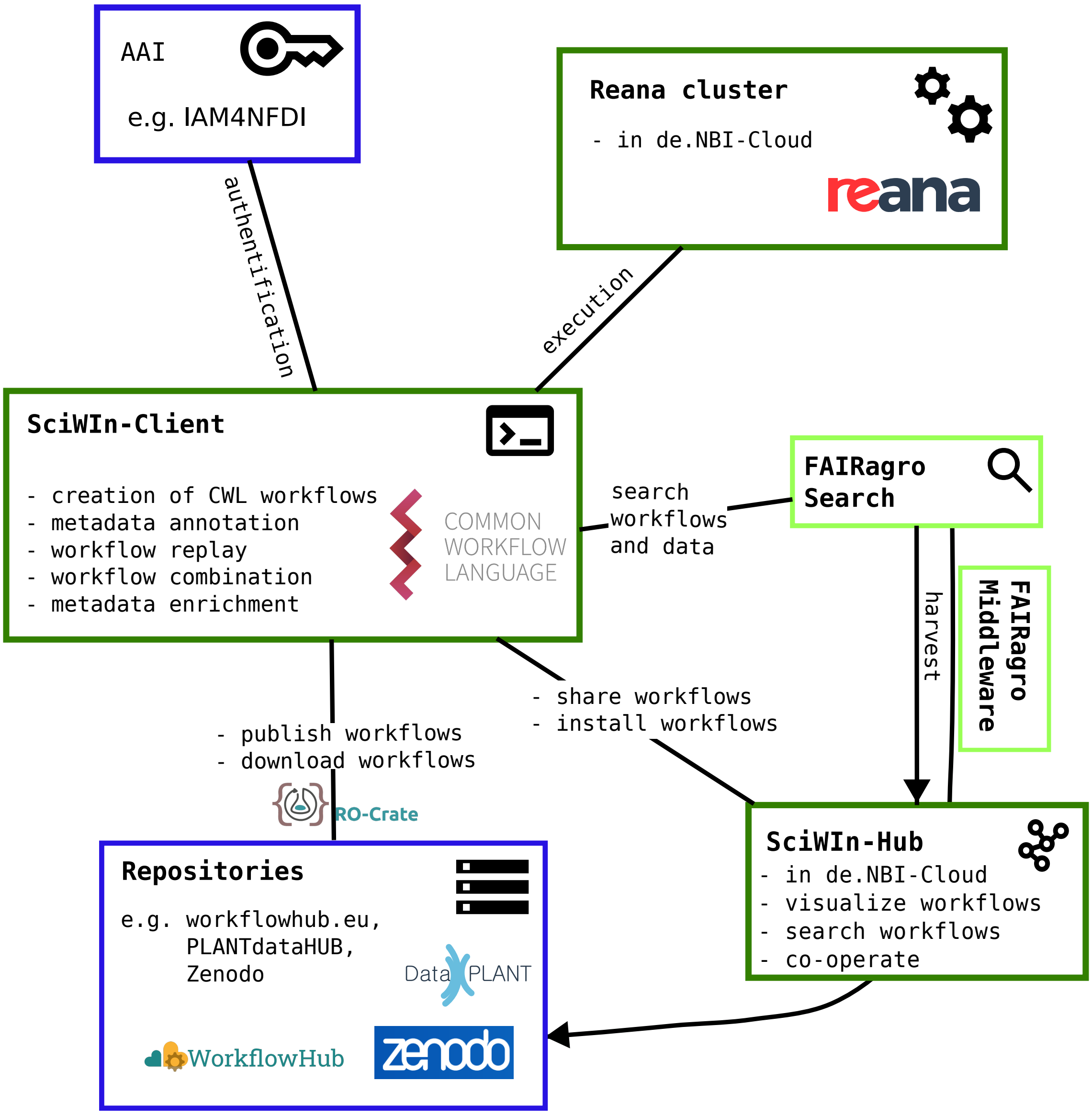

Figure 2: New ecosystem sketch

The main areas of sctivity of Measure 4.4 / SciWIn can therefore be listed as follows:

1. The SciWIn client

- effortless recording of computational workflows in CWL
- remote execution of computational workflows
- local management of multiple workflows
- authorization against an AAI

2. SciWIn-Hub

- sharing of computational workflows
- collaboration platform for computational work
- make workflows available for harvesting by FAIRagro-Search / Middleware

settled on the workflow description language CWL (Crusoe et al. (2022)), on - Compute instances: in principle general CWL capable, in practice reana, in principle any Reana in practice our own - Workflow objects : Move from ARC as central data structure to multi-data container compatibility.

New components:

AAI Client Hub Storage Instances Workflow objects now multi-paradigm compatibility based on provenance run crate.

SciWIn client

- requirements
- target group
- tech stack
- use cases

conceptual challenges

remote execution Reana wegen Multicloud workflow & data remote reference dockerization

Selection of CWL

Selection of

SciWIn-Hub

In progress + requirements + target group + tech stack + use cases

- "CWL package manager"
- Visualisierung von Workflows
- cordra …

Development strategy

Interaction within FAIRagro

- helpdesk
- use cases
- outreach / promotion
- techn : search findet workflows

Interaction within NFDI and beyond

- BASE4NFDI (KG4NFDI, MC (Reana), IAM4NFDI)

Towards the establishment of SciWIn as a common tool for computational workflows

Use cases, Outreach Workshop, Identify other consortia to co-operate reach out to other communities

[diepenbroek2023] The concept of data integration is partly described in the RDC mediation layer and covered by semantic tools (Glöckner et al., 2020), and a concept for integrated data and process storage is part of the DataPLANTs ARC model (Krantz et al., 2021)

Overall structure: SciWIn-Client and SciWIn-Hub

- A client-part that works de-centralized, independent of any central infrastructure as part of the scientists' daily toolset.
- Serves to capture computational workflows with minimal effort in a standards compliant way.

Quotes to be mined

Develop the concept of SciWIn (Measure 4.4) jointly together under the umbrella of the NFDI-RDC and collaborate with NFDI4BioDiversity and DataPLANT as well as the prototypic container deployment to the de.NBI cloud node at BLU;

A SciWIn pilot will be rolled out at de.NBI operated by Bielefeld University (BLU).

The concept of data integration is partly described in the RDC mediation layer and covered by semantic tools (Glöckner et al., 2020), and a concept for integrated data and process storage is part of the DataPLANTs ARC model (Krantz et al., 2021), but a fully integrated infrastructure is provided by neither of those consortia. Therefore, Measure 4.4 will extend the two architectural designs and provide a workflow infrastructure that applies the FAIR DO concepts (Measure 3.5), the service middleware components (Measure 4.2) and its own workflow hub as an easy-to-use interface to work on and create new FAIR DO outputs with automatically annotated provenance graphs.

SciWIn will be part of the NFDI cross-cutting topic "RDC implementations", and therefore, its concept will be developed as a joint effort between FAIRagro, FAIR-DS, Dataplant and the NFDI section, RDC [as part of cross-cutting topics (Ebert et al., 2021)]. Coordinated in Measure 5.3, the SciWIn working group brings these stakeholders together and will be initiated by a kickoff meeting (M4.4.1). Thus, the fundamental principles, architectures and interfaces are described, and a coordinated concept will be created and published (D4.4.1), which will incorporate the RDC ideas and DataPLANTs ARC model.

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

De Smedt, K., Koureas, D., & Wittenburg, P. (2020). FAIR Digital Objects for Science: From Data Pieces to Actionable Knowledge Units. *Publications*, *8*(2), 21. https://doi.org/10.3390/publications8020021

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

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