3rd FAIR Digital Objects Conference

Extended Abstracts

https://doi.org/10.52825/xxxx........ DOI placeholder (WILL BE FILLED IN BY TIB Open Publishing)

Published: (WILL BE FILLED IN BY TIB Open Publishing)

Digital Shadows to Orchestrate Cross Institutional, Souvereign, Multiomics Dataflows in Plant Research

Theodor Strauch[1,*] https://orcid.org/0009-0003-4327-6437, Thomas Altmann[1] https://orcid.org/0000-0002-3759-360X, Daniel Arend[1] https://orcid.org/0000-0002-2455-5938, Jorge Garcia Brizuela[1] https://orcid.org/0000-0003-0527-1339, Antonia Leidel[1] https://orcid.org/0009-0007-1765-0527, Lena Perzlmaier[2] https://orcid.org/0000-0002-0487-2151, Elena Rey-Mazón[1] https://orcid.org/0000-0003-4813-5927, Uwe Scholz[1] https://orcid.org/0000-0001-6113-3518, Danuta Schüler[1] https://orcid.org/0000-0003-4277-9879, Claus Weiland[2] https://orcid.org/0000-0003-0351-6523, and Matthias Lange[1] https://orcid.org/0000-0002-4316-078X

[1]Leibniz Institute of Plant Genetics and Crop Plant Research, Gatersleben, Germany

[2]Senckenberg –Leibniz Institution for Biodiversity and Earth System Research, Frankfurt, Germany

*Correspondence: Theodor Strauch, strauch@ipk-gatersleben.de

Abstract. In this presentation we showcase a conceptual basis for FAIR by design principle for data driven research projects with a highly federated set-up, diverse technology readiness levels and high sovereignty of the partner. This is carried by a review of RDM practices in collaborative research projects over the last decade where the Leibniz Institute of Plant Genetics and Crop Plant Research (IPK) was in charge of data management workpackages. We will present hurdles and challenges, infer lessons learned and demonstrate a framework that makes use of the FAIR Digital Object (FDO) concept in a self-service infrastructure driven by strict data flow engineering and an open-source collaborative data space infrastructure.

Keywords: Annotated research Context, Data Flow, FAIR Digital Objects, Collaborative Data Space

1. Learning From Experience in Data Management

Research is conducted in a collaborative manner across organisational boundaries and domains. This enables the federation of experts, experimental facilities, technologies and resources in a beneficial way. This common set-up for scientific projects is on the same hand challenging for an efficient data life cycle to manage and preserve data and streamline the creation of interoperable and consistent data sets toward FAIR Digital Objects. The challenges arise from the sovereignty of research units with heterogeneous data management policies, the heterogeneity of research domains with diverse data types, scales, formats and metadata standards. This complexity is amplified further in public-private research partnerships, where the partners have diverse governance systems, data management and IT readiness levels.

A reference for a multinational data science project is the Activated GEnebank NeTwork (AGENT) [2] with 18 European genebank and plant science research centers involved that showed a broad range of technology and data management readiness levels. The data management objective was to activate historic Wheat and Barley field observation data, enrich them with modern genotyping data and make plant material digitally accessible for investigation. A strictly monitored data flow, encompassing rule-based metadata and data validation pipeline [3], a collaborative data curation and a centrally managed database infrastructure [4] was identified as a beneficial approach. The German national public-private partnership project

Advanced Virtual and Augmented Reality Approaches in Seeds to Seeds (AVATARS) [5] and German National research data Infrastructure (NFDI) consortium for a FAIR Data Infrastructure for Agrosystems (FAIRagro) [6, 7], based on a similar institutional readiness level but were characterised by multi-omics domain and diverse governance models in private and public partners. In AVATARS, a strict top-down digital shadow based data flow across partner data infrastructures was devised, resulting in a network of interlinked multi-omics datasets with complete sample and material provenance, enabling deep genotype x phenotype x environment modeling in a wide range of types and scales, reaching from tabular observation to high throughput imaging and genotyping data. In contrast, FAIRagro was faced with long-standing established institutional data management processes and research data infrastructures with a high degree of sovereignty and diverse technical readiness levels. This required the implementation of a middleware [8], which used the smallest common access path by harvesting metadata from the exposed Web landing pages of the data sets or existing RESTful Services of the particular repositories. As a result, an overarching data catalog across data sets residing in autonomous data repositories is aggregated as JSON files and indexed by the FAIRagro Search Hub [9]. For metadata harmonisation the retrieved metadata is mapped to AgriSchema [10], an extension to schema.org [11], first by repository specific mapping rules.

2. **Leveraging the FDO-Concept**

Experience gathered in projects and consortia that are characterised by a high degree of diversity in governmental structures and policies as well as a broad range of readiness levels on one side, and high demands on FAIR data management on the other side can be distilled into four core lessons learned: (a) data will be increasingly heterogeneous and voluminous and requires data type agnostic collaborative data space and FDO representation, (b) significant effort is needed for data curation alongside each project, (c) complex, cross institutional data flows need to be orchestrated supported by a meticulous ID and metadata management, and (d) data and metadata needs to be persisted beyond a project's lifespan using community established, data domain agnostic infrastructures that are operated and financed within institutional service portfolio.

Because of inherent diversity in research projects, top-down approaches that force and finance FAIR by design data management are unlikely to resolve these problems. We are convinced that the FDO concept underpinned by a strictly monitored data flow and distributed, collaborative data space, equipped with a tooling ecosystem could help to overcome these challenges. This conceptual principle is currently being implemented in recently granted research projects, like INTEGRA, a public-private partnership to enhance genetic gain towards resilient rape seed, and the German NFDI consortia DataPlant and FAIRagro, building national plant and agronomic research data infrastructures. Here, one of the technical pillars is a collaborative data space in the form of a GitLab based PLANTdataHub [12] linked to a data store backend using GIT-LSF. This Annotated Research Context (ARC) concept addresses challenge (b) by facilitating domain scientists, data management experts and data scientists to work together in one common data space, that is technologically supported by CI/CD pipelines for regularly scheduled validation of metadata quality [3] and an RO-CRATE artifact build pipeline to release and publish FDOs at domain accepted repositories, like e!DAL-PGP [13]. In this talk we will showcase by the example of the INTEGRA project, how the Minimum Information About Plant Phenotyping Experiments (MIAPPE) [14] standard mapped into the ARC's ISA-MIAPPE serialisation [15] will be used to harmonize heterogeneous data from different sources, partially addressing challenge (a). Furthermore, we will show the integration of a PLANTdataHub instance to an object store in the form of a Ceph-S3 Gateway for Git-LFS to meet the expected data volume requirements, fully addressing challenge (a). Facing challenge (c), a lightweight web-based data registry will be outlined that is the front end for INTEGRA's data flow specification, which enables project partners to register materials, samples, experiment data with a project internal ID and initial metadata. These project scoped IDs will then be used to construct ARCs as they evolve throughout the project's lifetime, to integrate material and data flow towards a data mesh of qualified connected data artifacts. These project IDs will be subsequently used to mint permanent unique identifiers (PUID) in the release and publication process, e.g. IGSN for samples, DOI for material and FDO itself. The resulting ARCs are self-contained, metadata annotated and versioned data units that address challenge (d) by enabling streamlined release as DOI citable data publications, while incurring minimal data curation effort. (Figure 1) In addition to the classical Handle-based FDO approach, ARCs can also be published as webby FDOs by using their RO-Crate representation in combination with FAIR Signposting, enhancing the machine-actionable discoverability and accessibility [16,17].

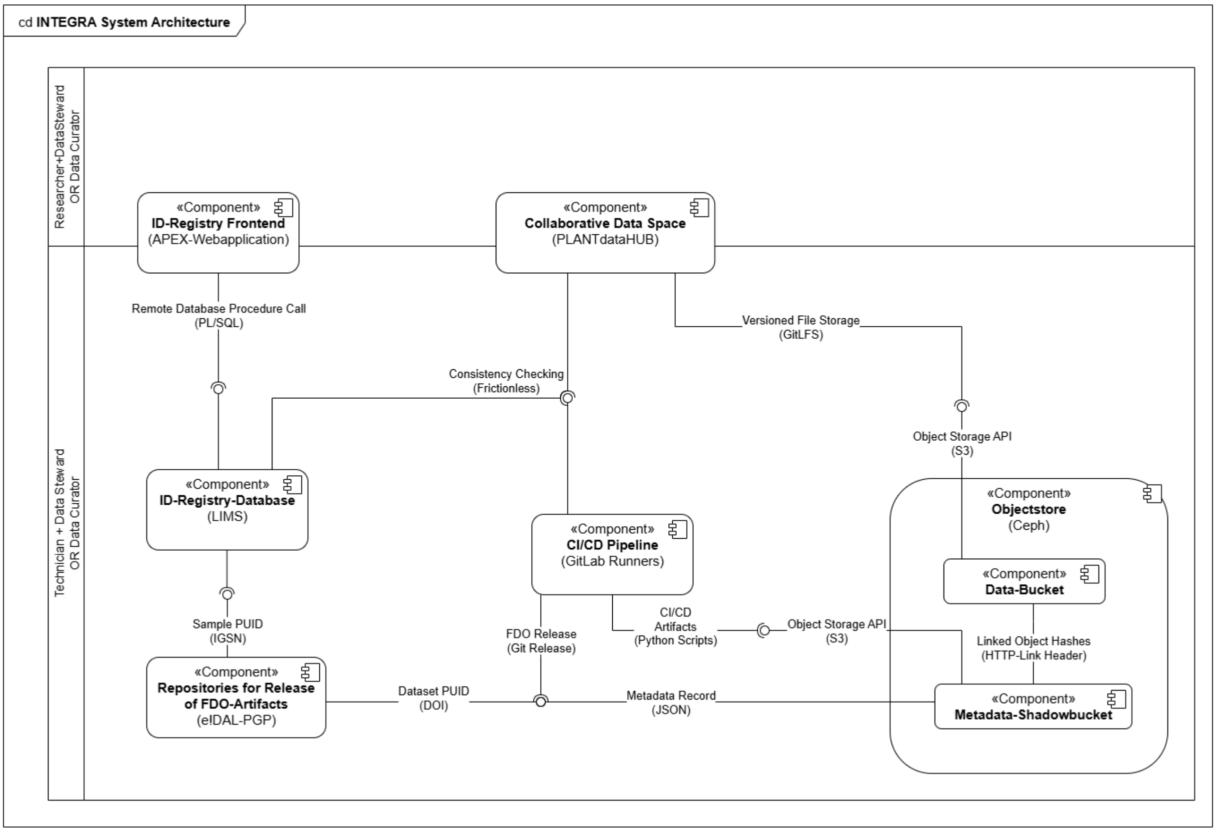

Figure 1: Conceptual system architecture of INTEGRA dataspace is represented as a UML diagram, wherein the generic framework components are represented as bold text and the concrete implementations in INTEGRA are represented as text in brackets. Researchers and technicians, accompanied by a data steward (self-service project set-up) or a data curator (full-staffed project service) initially register used biological materials, collected samples and obtained results through a Web-Frontend to obtain consistent identifier in project scope, e.g. INTEGRA-IDs. These IDs are then used in a PLANTdataHub instance, serving as a collaborative data management environment, to feature a consistent data provenance, enabling the construction of a data mesh to build validated and consistent FDO-artifacts through CI/CD-pipelines. This will then facilitate the streamlined publication of data and metadata into FAIR domain repositories.

Data availability statement

Data from the FAIRAgro consortium is available in the FAIRAgro Search Hub at https://search-hub.fairagro.net. Data from the AGENT project is available in the AGENT-Webportal at https://agent.ipk-gatersleben.de. Data from the AVATARS project is available to authenticated and authorized members only at https://apex.ipk-gatersleben.de/apex/f?p=avatars. Data from the DataPlant consortium is available in the PlantdataHub at https://git.nfdi4plants.org/explore. Due to the project being in an early conceptual phase, data from the INTEGRA project is not yet available.

Author contributions

Theodor Strauch: Writing - Original Draft, Methodology, Software, Visualization; Thomas Altmann: Funding acquisition, Project administration; Daniel Arend: Writing - Original Draft, Conceptualization, Methodology; Jorge Garcia Brizuela: Software, Methodology; Antonia Leidel: Software; Methodology; Lena Perzlmaier: Writing - Review & Editing; Elena Rey-Mazón: Data Curation, Validation; Uwe Scholz: Funding acquisition; Danuta Schüler: Data Curation, Software, Validation; Claus Weiland: Writing - Original Draft; Matthias Lange: Writing - Original Draft, Supervision, Funding acquisition

Competing interests

The authors declare that they have no competing interests.

Funding

The work was supported by the German Research Foundation DFG under the grant agreement number 501899475 (FAIRAgro) and by the Federal Ministry of Research, Technology and Space under the grant agreement number 031B1548A (INTEGRA).

References

[1] Schüler D, Lange M, Altmann T, et al (2025) Data management in balance – a decade of balancing pragmatism, sustainability and innovation at plant research center IPK Gatersleben. J Integr Bioinforma 20250012. https://doi.org/10.1515/jib-2025-0012
[2] Global Access to Plant Genetic Resources | AGENT. https://agent-project.eu/. Accessed 23 Oct 2025
[3] Fowler D, Barratt J, Walsh P (2018) Frictionless Data: Making Research Data Quality Visible. Int J Digit Curation 12:274–285. https://doi.org/10.2218/ijdc.v12i2.577
[4] Alaux M, Adam-Blondon A-F, Brouwer M, et al (2024) AGENT Guidelines for dataflow
[5] AVATARS – Advanced Virtuality and AugmenTed Reality AppRoaches in Seeds to Seeds. https://www.avatars-project.de/. Accessed 23 Oct 2025
[6] Ewert F, Arend D, Asseng S, et al (2023) FAIRagro FAIR Data Infrastructure for Agrosystems - Proposal 2021. https://doi.org/10.5281/ZENODO.7528172
[7] Specka X, Martini D, Weiland C, et al (2023) FAIRagro: Ein Konsortium in der Nationalen Forschungsdateninfrastruktur (NFDI) für Forschungsdaten in der Agrosystemforschung: Herausforderungen und Lösungsansätze für den Aufbau einer FAIRen Forschungsdateninfrastruktur. Inform Spektrum 46:24–35. https://doi.org/10.1007/s00287-022-01520-w
[8] García Brizuela J, Scharfenberg C, Scheuner C, et al (2024) A roadmap for a middleware as a federation service for integrative data retrieval of agricultural data. J Integr Bioinforma 21:20240027. https://doi.org/10.1515/jib-2024-0027
[9] FAIRagro Search Hub. https://search-hub.fairagro.net/. Accessed 23 Oct 2025
[10] Arend D, Beier S, Bouri L, et al (2022) BioHackEU22 Project 22: Plant data exchange and standard interoperability
[11] Castro LJ, Fluck J, Arend D, et al (2023) Schema.org as a Lightweight Harmonization Approach for NFDI. https://doi.org/10.5281/ZENODO.8331237
[12] Weil HL, Schneider K, Tschöpe M, et al (2023) PLANTdataHUB : a collaborative platform for continuous FAIR data sharing in plant research. Plant J 116:974–988. https://doi.org/10.1111/tpj.16474
[13] Arend D, König P, Junker A, et al (2020) The on-premise data sharing infrastructure e!DAL: Foster FAIR data for faster data acquisition. GigaScience 9:giaa107. https://doi.org/10.1093/gigascience/giaa107
[14] Papoutsoglou EA, Faria D, Arend D, et al (2020) Enabling reusability of plant phenomic datasets with MIAPPE 1.1. New Phytol nph.16544. https://doi.org/10.1111/nph.16544

[15] Sansone S-A, Rocca-Serra P, Gonzalez-Beltran A, et al (2016) Isa Model And Serialization Specifications 1.0. https://doi.org/10.5281/ZENODO.163640

[16] Soiland-Reyes S, Sefton P, Leo S, et al (2025) Practical webby FDOs With RO-Crate and FAIR Signposting: Experiences and Lessons Learned. Open Conference Proceedings, 5. https://doi.org/10.52825/ocp.v5i.1273

[17] Weiland C, Grieb J, Bauer D, Chala D, Kusch E, Andrew C, Endresen D (2024) Dataspace Integration for Agrobiodiversity Digital Twins with RO-Crate. Biodiversity Information Science and Standards 8: e134479. https://doi.org/10.3897/biss.8.134479