ARC as a FAIR Digital Object: Declarative Provenance Annotation for Scalable Research Data Management

The rapid growth of high-throughput and multi-modal measurement technologies has led to an unprecedented proliferation of heterogeneous research data. Contemporary experiments often integrate diverse analytical and computational methods, producing data that differ widely in format, structure, and semantics. To render such data interoperable, and reusable over time, it is essential to capture not only the data itself but also the provenance of the actions that generated it. The Annotated Research Context (ARC) addresses this challenge by defining a comprehensive FAIR Digital Object (FDO) specification for representing complete research cycles through declarative process annotation.

ARC extends established models, namely the Investigation Study Assay (ISA) framework for experimental descriptions and the Common Workflow Language (CWL) for computational workflows. Both models share a procedural approach, in which datasets are the products of sequential processes. ARC generalizes this approach, providing a process-centric annotation model that spans both the wet-lab and dry-lab portion of a research project. This allows for reuse of annotations especially in multimodal experiments, where prepared samples are measured using different technologies. All data resulting from these measurements by definition inherit the traits of the source material, in addition to the traits describing the processes producing them.(Rocca-Serra et al. 2012)

This generic, process-based model can be infused with domain specific knowledge through semantics in a key–key–value structure. The first key is the property of the annotation model and can be interpreted as a term from one of the two *container ontologies* (ISA or CWL). It assigns the following key-value pair a structural role, such as a process parameter, material attribute, or a software component. The second key originates from a *domain ontology* that specifies the scientific or technical concept (e.g. “mass spectrometer” or “growth medium”). Lastly, the value describes the concrete instance of this entity in the experiment, and is either selected from the domain ontology, or a numeral. This flexible annotation principle enables the processes to be used in highly differing domains and allows the ARCs of those domains to grow and evolve alongside the ontologies, following the advancements and improvements in those domains.

To ensure interoperability and machine actionability, ARCs can be represented as RO-Crates, a linked-data influenced research packaging standard which is quickly gaining traction across the European research data landscape. RO-Crate provides a consistent metadata structure, laying a foundational guideline for data file storage and annotation. On top of this, it is easily extendable using profiles, in our case the ARC RO-Crate profile, formalizing the aforementioned annotation model into a JSON-LD representation. On the other hand, for direct engagement by experimental researchers, a complementary representation known as the ARC Scaffold is provided. The ARC Scaffold adopts a tabular, spreadsheet-based format based on the ISA-Tab format, that simplifies metadata entry while preserving semantic richness. Bidirectional conversion between ARC RO-Crates and ARC Scaffolds is fully supported, allowing researchers to transition seamlessly between user-friendly and machine-readable formats without loss of information.(Florian Wetzels et al. 2025; DataPLANT 2025; Soiland-Reyes et al. 2022)

The ARC specification is supported by an extensive ecosystem of tools that promote usability, integration, and sustainability. The core ARCtrl software library is implemented in .NET, Python, and JavaScript and provides programmatic interfaces for managing ARC content stemming from both Scaffold and RO-Crate form. Through a thorough data model and exhaustive functionality, it supports easy implementation of higher-level tools, catalyzing user benefit. One central higher-level tool is SWATE, a web application which allows ontology-driven metadata creation directly within familiar spreadsheet environments, lowering the barrier for domain scientists to produce FDO-compliant annotations. Besides being usable directly as a plugin in Excel, SWATE is also the main building block in the ARCitect desktop application. The ARCitect is a one-in all interface for visually handling and modifying ARCs, including Git-based local versioning functionality in addition to integrating ARCtrl and SWATE for providing easy ARC-level IO and metadata annotation. (Weil et al. 2025)

The core toolstack is complemented by the PLANTdataHUB, a GitLab-based collaborative environment, which provides a platform for version-controlled collaborative management of ARC projects. By leveraging Git for tracking contributions and metadata evolution, it adds an additional dimension of provenance, realizing the principle of “immutable yet evolving” data: while archived versions remain persistent, annotations can be refined collaboratively over time. The resulting workflow replaces unstructured communication with transparent, traceable metadata evolution, reinforcing the STAP trust dimensions of data sovereignty, transparency, accountability, and persistence.(Weil et al. 2023)

Publication bibliography

DataPLANT (2025): nfdi4plants/ARC-specification: Annotated Research Context Specification v3.0-draft.2: Zenodo.

Florian Wetzels; Kevin Schneider; Stuart Owen; Lukas Weil; DataPLANT (2025): nfdi4plants/isa-ro-crate-profile: ISA RO-Crate Profile v1.0-draft.1: Zenodo.

Rocca-Serra, Philippe; Maguire, Eamonn; Taylor, Chris; Field, Dawn; Wittenberger, Timo; Santarsiero, Annapaola et al. (2012): Investigation-Study-Assay, a toolkit for standardizing data capture and sharing. In Lee Harland, Mark Forster (Eds.): Open source software in life science research. Practical solutions to common challenges in the pharmaceutical industry and beyond. Philadelphia: Woodhead Pub (Woodhead Publishing series in biomedicine, no. 16), pp. 173–188. Available online at https://www.sciencedirect.com/science/article/pii/B9781907568978500072.

Soiland-Reyes, Stian; Sefton, Peter; Crosas, Mercè; Castro, Leyla Jael; Coppens, Frederik; Fernández, José M. et al. (2022): Packaging research artefacts with RO-Crate. In *Data Science* 5 (2), pp. 97–138. DOI: 10.3233/DS-210053.

Weil, Heinrich Lukas; Frey, Kevin; Schneider, Kevin; Ott, Caroline; Wetzels, Florian; Mühlhaus, Timo (2025): nfdi4plants/ARCtrl: 3.0.0-beta.10: Zenodo.

Weil, Heinrich Lukas; Schneider, Kevin; Tschöpe, Marcel; Bauer, Jonathan; Maus, Oliver; Frey, Kevin et al. (2023): PLANTdataHUB: a collaborative platform for continuous FAIR data sharing in plant research. In *The Plant journal : for cell and molecular biology* 116 (4), pp. 974–988. DOI: 10.1111/tpj.16474.