



Report on Metadata & PID Workflows

# Integration of Persistent Identifiers into ELN and DMP Tools for NFDI

# Imprint

**Authored by** Sara El-Gebali  
(<https://orcid.org/0000-0003-1378-5495>)

Antonia C. Schrader  
(<https://orcid.org/0000-0001-7080-634X>)

**Contributions by** Stephanie Hagemann-Wilholt  
(<https://orcid.org/0000-0002-0474-2410>)

**Date of publication** April 20, 2026

**Version** Version 1.0

**DOI** <https://doi.org/10.5281/zenodo.19607306>

**License:** This work is licensed under  
<https://creativecommons.org/licenses/by/4.0/>

## **Declaration of AI Usage:**

This report was drafted with the assistance of generative artificial intelligence for the purposes of language translation, stylistic review, and editorial refinement. All final outputs were manually verified and curated by the project team to ensure technical accuracy and consistency.

## **About**

PID4NFDI (<https://base4nfdi.de/projects/pid4nfdi>) is the basic service for persistent identifiers in development for the German National Research Data Infrastructure (NFDI). PID4NFDI is part of and funded through Base4NFDI.

Funded by DFG as part of NFDI. DFG Grant Number: 521453681

# Content

1. Introduction.....	3
2. About the PID Coordination Hub.....	3
3. Why NFDI Consortia need Persistent Identifiers within DMPs and ELNs.....	4
4. Conceptual Framework: Data Lineage and Interoperability.....	5
4.1 Defining Data Lineage in the Research Context.....	5
4.2 Technical Interoperability through Standardized Identifiers & Metadata Schema.....	7
4.3 Crosswalk.....	8
5. Building the Community of Practice.....	9
5.1 Approach and Participation.....	9
5.2 DMP Focus Group: From static Planning to machine-actionable Metadata.....	10
5.3 ELN Focus Group: PIDs as active Research Tools.....	11
5.4 Joint Workshop, Berlin: From 22 to 23 September 2025.....	13
Day 1: Interoperability in Practice.....	14
ELN Stream.....	14
DMP Stream.....	15
Day 2: Standards, Quality, and Implementation.....	15
6. The Next Phase: Incubator Projects with TS4NFDI.....	18
6.1 From Community Alignment to Technical Prototyping.....	18
6.2 Incubator Project 1: DataCite Terminologies in TS4NFDI.....	19
6.3 Incubator Project 2: Semantic ELN Integration (RSpace).....	19
6.4 Timeline, Sustainability, and what comes Next.....	20
References.....	21

## Abstract

This report documents the PID Coordination Hub's coordinated effort to integrate persistent identifiers (PIDs) into the research data lifecycle through data management plans (DMPs) and electronic lab notebooks (ELNs). During 2025, the endeavor included establishing two focus groups. Each group held three virtual meetings from spring through summer, culminating in a joint in-person workshop in Berlin in September 2025. The report summarizes findings on metadata workflows, PID integration pathways, and technical interoperability requirements. These findings establish a foundation for two incubator projects that will launch in 2026 in collaboration with TS4NFDI. The projects aim to operationalize the DataCite metadata schema as a canonical reference model within the NFDI and demonstrate how capturing high-quality, standards-aligned metadata early on can enable seamless PID workflows across the research data lifecycle.

# 1. Introduction

This document outlines the strategic framework and findings of the PID4NFDI Coordination Hub regarding the integration of Persistent Identifiers (PIDs) into Data Management Plans (DMPs) and Electronic Lab Notebooks (ELNs).

During the Integration Phase of PID4NFDI two research data lifecycle-orientated focus groups were set up to evaluate current DMP and ELN features and identify requirements for holistic PID integration.

The document synthesizes the outcomes of the Focus Groups meetings in 2025 and the subsequent in-person workshop. It describes the role of DMPs and ELNs for PID workflows and how they contribute to interoperability.

The results of these community activities will be continued in an incubator project with the [Terminology Services 4 NFDI](#) (TS4NFDI) in 2026. The roadmap for these efforts is also outlined here.

## 2. About the PID Coordination Hub

The PID Coordination Hub, established under the Base4NFDI-funded project PID4NFDI, serves as the central point of coordination for PID activities within the National Research Data Infrastructure (NFDI). Its core mandate is to promote consistent and interoperable PID practices across all NFDI consortia, ensuring that research outputs, actors, and infrastructures are globally discoverable and persistently linked. The Hub's mission is to align PID services, metadata standards, and implementation strategies across domains, creating a shared foundation for interoperability and long-term data integrity.

Working in close collaboration with NFDI stakeholders including the project partners DataCite, the Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen (GWDG), the TIB – Leibniz Information Centre for Science and Technology, and the Helmholtz Open Science Office, the Coordination Hub develops practical guidance, shared tools, and support mechanisms that ensure alignment between NFDI and the international PID landscape.

### 3. Why NFDI Consortia need Persistent Identifiers within DMPs and ELNs

In the highly diverse landscape of the NFDI, where consortia represent vastly different scientific disciplines and unique data cultures, the strategic integration of Persistent Identifiers (PIDs) into DMPs and ELNs serves as a critical unifying bridge. While domain-specific requirements often lead to fragmented toolsets and siloed metadata practices, PIDs provide a universal anchor that shifts the administrative burden from a manual chore to a reusable asset. By capturing identifiers for people (e.g. ORCID), organizations and funders (such as ROR), and projects (such as RAiD) at the earliest possible stage, researchers can overcome the recurring challenge of re-entering the same information across disparate systems.

This **"enter once, reuse everywhere"** approach ensures that high-quality, standardized information flows seamlessly from the initial planning phase in a DMP to specialized experimental workflows in an ELN, and finally into cross-disciplinary repositories. To fully realize these benefits, however, the NFDI must move beyond traditional PID usage, like DOIs for (text-)publications and focus on **early adoption** within the research data lifecycle. Our landscape analysis<sup>1</sup> has identified significant fragmentation of the PID landscape within NFDI consortia, marked by diverse application scenarios and varying levels of PID integration maturity. Addressing this fragmentation requires integrating PIDs into the very tools where research is planned and executed: **Data Management Plans** and **Electronic Lab Notebooks**.

- **In DMPs**, e.g. the integration of PIDs for projects (e.g., RAiDs) and awards enables standardized selection and machine-processable workflows. This allows consortia to harmonize metadata across disciplines and track project impacts from the outset.
- **In ELNs**, e.g. PIDs for instruments, methods, and material samples streamline workflows by documenting the research process in real-time. This ensures that highly granular entities are linked to their contextual details long before reaching the publication stage.

---

<sup>1</sup> El-Gebali, S., & Böhm, J. (2025). Landscape Analysis of PID Practices in NFDI (1.2). Zenodo. <https://doi.org/10.5281/zenodo.15689799>.

Ultimately, by coherently embedding PIDs into DMPs and ELNs, the NFDI establishes a **"FAIR-by-design"** framework. This proactive approach significantly reduces redundant data entry for the individual researcher while ensuring that research outputs remain visible and trackable for the community. By improving technical interoperability with global infrastructures like DataCite, ORCID, and EOSC, the NFDI not only meets its own strategic goals but also fulfills the rigorous expectations of funders and the broader international research community.

## 4. Conceptual Framework: Data Lineage and Interoperability

At the heart of this initiative lies the concept of data lineage, the documented history of how data moves through the research lifecycle, from initial creation through analysis, publication, and reuse<sup>2</sup>. Data lineage is essential for reproducibility, attribution, and impact assessment. When metadata and identifiers are captured early in research workflows and propagated consistently across systems, data lineage becomes traceable, machine-readable, and ultimately reusable.

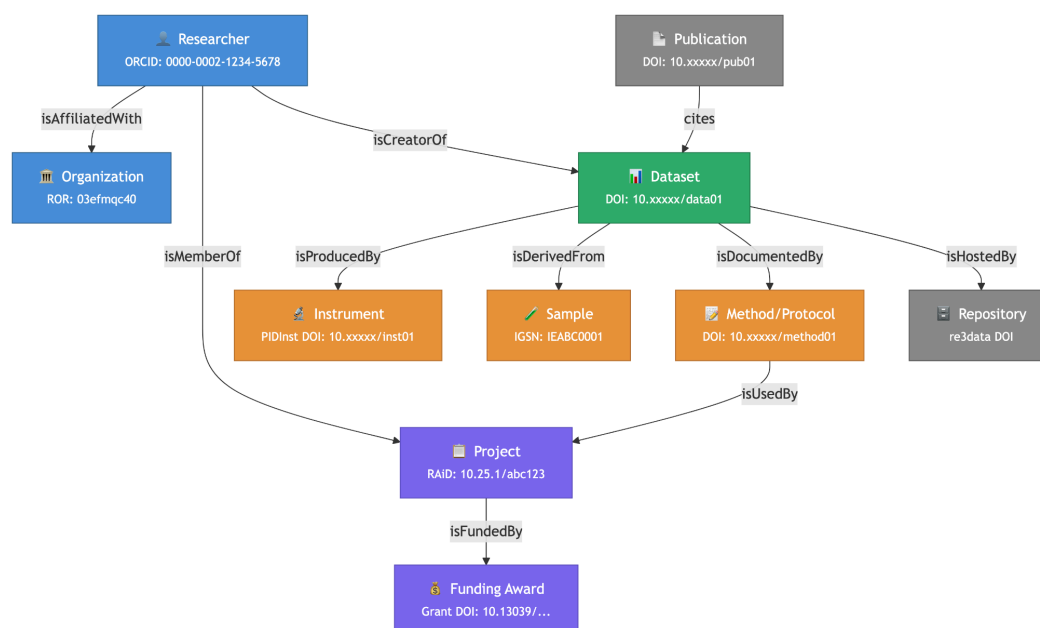
For the NFDI, establishing data lineage requires both technical and organizational mechanisms. Technically, PIDs provide stable anchors for connecting entities (people, organizations, datasets, instruments, projects) across systems. Organizationally, standard metadata schemas and controlled vocabularies enable disparate tools to interpret metadata consistently. When DMPs and ELNs capture metadata that aligns with these schemas and vocabularies, metadata flows smoothly from creation through repositories and to downstream discovery services.

### 4.1 Defining Data Lineage in the Research Context

Data lineage, as described above, answers questions such as: who created this dataset, with which instrument, under which project, using what method, and where was it deposited? When each of these entities carries a persistent identifier and the relationships between them are recorded in structured metadata, the lineage becomes not only human-readable but machine-actionable, enabling automated discovery, impact tracking, and reproducibility verification.

---

<sup>2</sup> What is data lineage? <https://www.ibm.com/think/topics/data-lineage>



**Figure 1.** The colour groupings reflect the lifecycle stages: blue for actors (**who**), purple for planning (**under which project/funding**), orange for execution (**with what instrument, sample, method**), green for the output (**dataset**), and grey for dissemination (**repository, publication**). Every edge is a typed relationship that would be recorded via DataCite relationType values, and every node carries a PID that makes it globally resolvable. That's the core point, the lineage is the graph itself, and the PIDs are what make it machine-actionable rather than just a narrative description.

The challenge is that current DMP and ELN tools operate largely in isolation, each with its own metadata structure. A researcher planning work in a DMP captures project details and dataset descriptions that never automatically reach the ELN where data is actually generated. Similarly, metadata created in the ELN, such as instrument configurations or sample origins, is not fed back to the DMP for documentation or passed on for publications in repositories. This fragmentation means that critical context is lost, metadata must be re-entered multiple times, and downstream repositories receive incomplete or inconsistent information.

For the NFDI, where [27 consortia](#) span disciplines from the humanities to the natural sciences, metadata fragmentation is not merely an inconvenience but a structural barrier to the interoperability and reuse that the infrastructure is designed to enable. Establishing data lineage therefore requires embedding PIDs and standardized metadata at the earliest points of the research process, in the DMP where work is planned and the ELN where it is executed, so that provenance information is captured once and propagated consistently downstream.

## 4.2 Technical Interoperability through Standardized Identifiers & Metadata Schema

PIDs and schema alignment serve as the bridge between isolated tools. Each PID type anchors a specific class of entity in the research ecosystem. However, identifiers alone are not sufficient. The metadata records that surround them must also be interoperable. The [Taskforce Metadata](#) within the NFDI Section [\(Meta\)data, Terminologies, Provenance](#) recommends three generic metadata schemas as the primary points of alignment: [DCAT](#), [schema.org](#), and [DataCite](#). Within this context, the DataCite metadata schema offers particular advantages as an initial canonical reference model. It is widely adopted globally (with more than 100 million DOIs registered), and provides a well-defined structure of mandatory, recommended, and optional fields. This structure allows DMP and ELN providers to begin with a clear set of core attributes while retaining flexibility for discipline-specific extensions. It is widely used within the NFDI<sup>3</sup>. This means that aligning DMP and ELN fields to DataCite attributes reduces the overall number of schema mappings that must be maintained while preserving the flexibility required across heterogeneous domains.

In practice, this means that when a researcher creates a dataset in an ELN, the tool can automatically populate metadata such as creator (ORCID), organization (ROR), instrument (PIDInst), and sample (IGSN) through lookup services, ensuring high data quality at the point of creation. When that dataset moves to a repository or PID service, the same metadata, now enriched and standardized, flows forward. The repository does not need to ask for the information again; it can trust and reuse the data already captured. This reduces burden on researchers, improves metadata consistency, and enables true end-to-end data lineage across the research lifecycle.

---

<sup>3</sup> El-Gebali, S., & Böhm, J. (2025). Landscape Analysis of PID Practices in NFDI (1.2). Zenodo. <https://doi.org/10.5281/zenodo.15689799>



## 4.3 Crosswalk

Even with a shared canonical reference model, real-world tools will continue to use their own internal schemas. The integration assessment conducted by PID4NFDI<sup>4</sup> across NFDI consortia confirms this diversity concretely. Platforms such as

- [RSpace](#) record samples, instruments, projects, and notebook entries using its own metadata fields;
- [Chemotion](#) captures reactions, molecules, and analysis datasets through specified element structures;
- [NOMAD](#) records its metadata schema as DataCite Metadata Schema 4.6;
- [Coscine](#) uses individually created metadata profiles described as AIMS (created by researchers);
- [NFDI4Cat](#) services employ DCAT and LinkML-based schemas; and
- [RDMO](#) (Research Data Management Organizer) captures DMP metadata through its own attribute-question structure.

Across the platforms surveyed, over a dozen distinct object types were identified, from datasets, samples, and instruments to reactions, molecules, protocols, and controlled vocabularies, each described differently depending on the tool. Of these, only a small number currently have documented mappings to DataCite, and most platforms reported no formal crosswalk between their internal schema and external standards.

The mechanism that enables communication between these heterogeneous schemas without requiring any single tool to abandon its design is the [crosswalk](#)<sup>5</sup>: a documented, structured mapping between the elements of one schema and the corresponding elements of another. The integration assessment makes the need for such crosswalk tangible, it shows precisely where mappings are missing, where PID integration is planned but not yet realised, and where metadata validation relies on tool-specific logic rather than shared standards.

---

<sup>4</sup> Data Collection & Mapping Baseline  
<https://docs.google.com/spreadsheets/d/1YS0ZC7dCFxN3DLzmR-qlyye5UIAT-k2ID2xidGxEr8/edit?usp=sharing>

<sup>5</sup> For more details please read: <https://pid.services.base4nfdi.de/blog/crosswalks/>

A crosswalk specifies, for example, that the DataCite property `creators[].name` corresponds to the RDMO attribute for "principal investigator" and to the RDA-DMP field `contributor.name`. It records not just that these fields are related, but the nature of the relationship (exact match, broader match, narrower match) and any transformations needed to preserve semantic meaning. When crosswalks are formalized in machine-readable formats such as [JSKOS](#) or [SSSOM](#), they can be consumed by APIs and used at runtime, a tool exporting metadata can query a mapping service to translate its internal field names into the target schema on the fly.

[Terminology services Suite \(TSS\)](#) provides the vocabulary layer that underpins these crosswalk. They ensure that when one system says "Dataset" and another says "Collection," the mapping between them is explicit and maintained centrally rather than embedded in code. The TS4NFDI Cocoda-based mapping service is designed to serve exactly this function: hosting, versioning, and exposing crosswalk mappings via a REST API, so that any NFDI tool can retrieve schema equivalences on demand. This approach allows each tool to retain its distinctive design while participating in a larger ecosystem of interoperable services.

## 5. Building the Community of Practice

### 5.1 Approach and Participation

The PID Coordination Hub established two dedicated **focus groups** to move beyond theoretical requirements toward concrete implementation pathways for integrating PIDs into **DMPs** and **ELNs**. These groups were designed to collaboratively explore a broad range of key topics: PID registration and integration into existing platforms, metadata quality and completeness, lifecycle integration challenges, evaluation of current training resources and development of new best practice documentation, and policy alignment including compliance with funder requirements. Rather than running DMP and ELN work as separate tracks, the programme was designed from the outset for convergence: three online meetings were held for each group between April and August 2025, culminating in a joint workshop in Berlin on 22 to 23 September 2025.

The two groups operated on different structural models reflecting the landscapes they addressed. The **DMP focus group** engaged directly with [DMP4NFDI](#), the NFDI's central service for data and software management plans, ensuring that integration

work was coordinated with the primary service provider rather than developed in parallel to it. The **ELN focus group** ran as an open platform, drawing developers from **RSpace**, **eLabFTW**, and **Chemotion** alongside infrastructure providers, NFDI consortia representatives, and researchers across multiple disciplines. This diversity was both necessary and intentional, given the decentralised ELN landscape and the absence of a single coordinating body.

## 5.2 DMP Focus Group: From static Planning to machine-actionable Metadata

The virtual consultation with DMP4NFDI concentrated on the transition from static DMP documents to **machine-actionable DMPs (maDMPs)**, plans that exchange structured metadata with other systems rather than serving only as human-readable records. Three online meetings in 2025 progressively refined what that transition would require.

The first meeting established shared goals and surfaced an immediate gap: NFDI consortia had concrete needs for integrating PIDs for non-traditional entities such as instruments and samples that the current **RDMO** framework did not yet address. The central organizing concept that emerged was a **push/pull** model for metadata flow. RDMO should pull existing PID-linked metadata, ORCID for researchers, ROR for organizations, RAiD for projects, and grant DOIs for awards, to auto-populate DMP fields, reducing manual entry and improving consistency. In the reverse direction, RDMO should **push** structured DMP metadata out to ELNs, repositories, and publishing platforms.

Subsequent meetings worked through the technical preconditions for this bidirectional flow. Participants prioritised building crosswalk between RDMO's attribute structure and both the DataCite schema and the [RDA Common Standard](#) for maDMPs. Future integration with **TS4NFDI's** mapping tools was identified as the mechanism for keeping those crosswalk maintained and machine-readable. The group also examined how versioning, how a DMP evolves as a project progresses, should be captured and communicated to downstream systems.

A critical constraint emerged before implementation could begin. The maDMP standard is undergoing a major revision, with a new version expected mid-2026. The existing RDMO DataCite export plugin, developed around 2020, is no longer maintained and reflects only a subset of the current DataCite schema; full implementation would require simultaneous updates to both the RDMO question

catalogue and the plugin. The DMP strand has therefore been deferred until the standard stabilises and DMP4NFDI capacity allows. The analytical foundation laid during the focus group, the push/pull framework, the crosswalk analysis, and the plugin audit, is preserved and will serve as the starting point when this work resumes.

## 5.3 ELN Focus Group: PIDs as active Research Tools

The **ELN focus group** addressed a more fundamental shift: moving PIDs from labels applied at publication to tools embedded in active research workflows. Three meetings from May through August 2025 brought together ELN developers, infrastructure providers, NFDI consortia members, and researchers to map how metadata is currently captured and where it breaks down.

The first meeting established the baseline. Participants from FAIRagro and NFDI4Cat articulated the case for early-stage PID integration most clearly: by the time data reaches a repository, too much context has already been lost. The group converged on a practical starting point, integrating ORCID and ROR directly into ELN researcher and organization profiles, ensuring that every experiment record carries persistent, resolvable identifiers from the moment of creation. The group also addressed the question of centralised versus decentralised PID infrastructure (DataCite versus ePIC Handles and dARK), arriving at a pragmatic consensus: the two approaches are complementary, and systems should support both, maintaining local mirrors where possible to ensure resilience independent of any particular central service.

Subsequent meetings mapped **concrete metadata workflows**. Three distinct interaction patterns were identified:

1. **importing** (pulling metadata from ORCID records, ROR entries, and grant DOIs to auto-populate experiment records with provenance information);
2. **minting** (registering IGSNs for material samples and DOIs for methods or protocols directly from the ELN interface); and
3. **exporting** (transforming ELN records into DataCite-compatible or [RO-Crate](#)-formatted outputs that preserve data lineage as records move downstream to repositories).

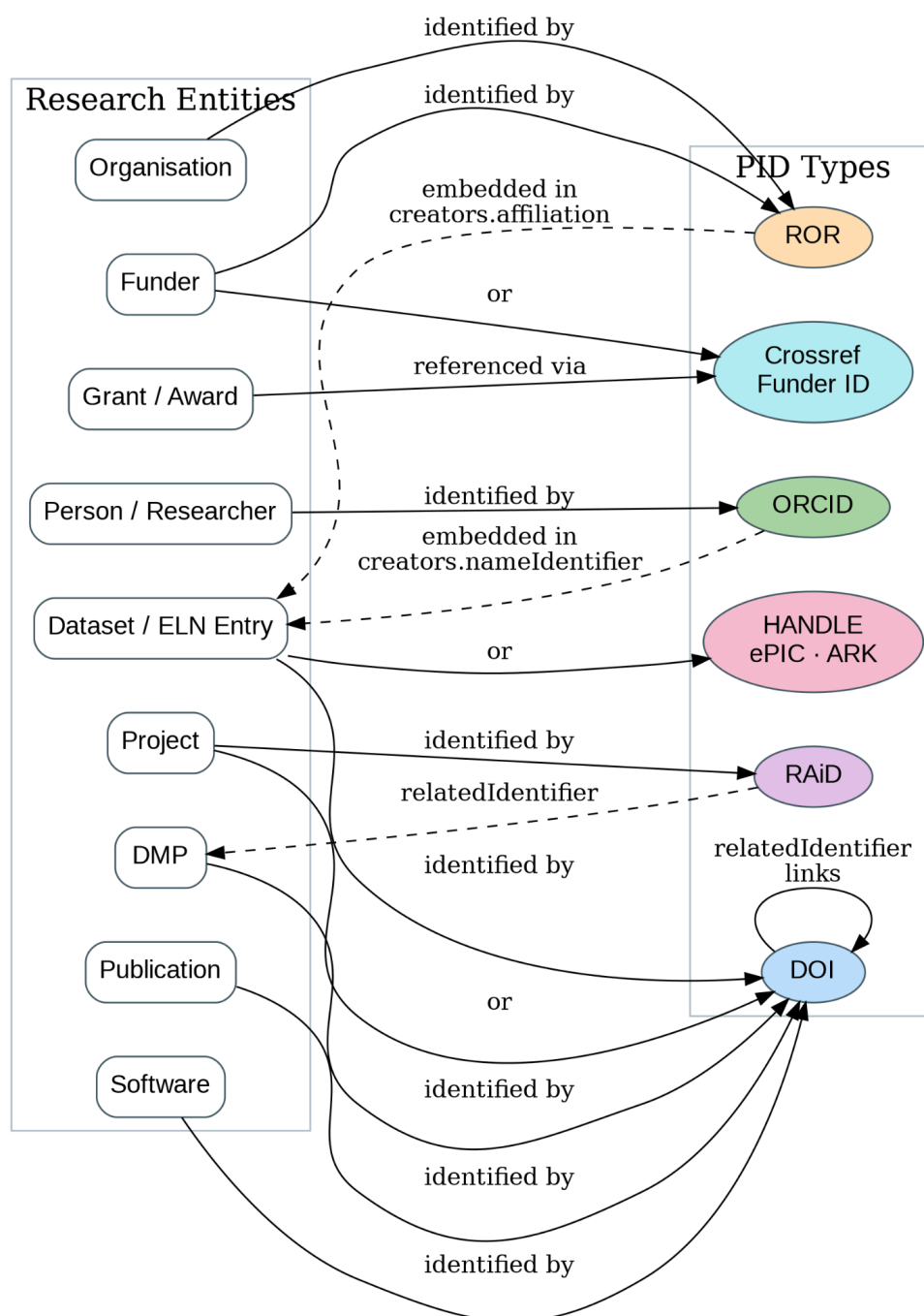
The final session focused on **PIDs for instruments (PIDInst)**. The group examined what makes an instrument a distinct citable entity, including the concept of an instrument as a data producer, and how to handle versioning when configuration

changes, and how to manage sensitive metadata such as equipment costs. A key finding emerged from examining global registries: instrument metadata is currently characterised by diversity and a lack of standardisation, with low compliance on optional fields<sup>6</sup> that carry the greatest benefit for reuse. The **relatedIdentifiers** field, which has the highest potential for connecting instruments to associated publications, datasets, and calibration records, is systematically underused and frequently misapplied. The infrastructure for richer instrument linkage already exists within the DataCite schema; the gap is adoption and correct use, not a missing capability. Addressing this will require standardised metadata templates, validation mechanisms, and tighter integration with ELN systems to capture instrument state and configuration changes systematically, the conditions for making instrument PIDs a viable part of the daily lab routine.

Three categories of actionable outcomes emerged from the ELN focus group, which successfully demonstrated that diverse tools and communities can converge on shared metadata principles while maintaining their distinctive designs and user experiences. Quick wins ready for near-term implementation include **ORCID auto-population** in user profiles, **ROR lookup** for organisation fields, and **controlled vocabulary widgets** for relation types. Persistent blockers requiring coordinated community effort include instrument versioning ambiguity, sensitive metadata handling, and variation in resource type definitions across disciplines. Shared architectural principles agreed on by all platforms are: early capture at point of entry, provenance-preserving export, and controlled-vocabulary selection rather than free text. A brief summary of all three meetings, including presentation materials and recommended readings, is publicly available on the **PID Coordination Hub website** in the Events section (<https://pid.services.base4nfdi.de/events/>).

---

<sup>6</sup> e.g. see optional metadata requirements within DataCite Metadata Schema: <https://datacite-metadata-schema.readthedocs.io/en/4.6/properties/overview/#levels-of-obligation>



**Figure 2.** The PID ecosystem: matching identifier types to research entities. A cross-cutting finding from both focus groups.

## 5.4 Joint Workshop, Berlin: From 22 to 23 September 2025

The consultative virtual phase culminated in an interactive in-person workshop hosted by PID4NFDI at the Helmholtz Association headquarters in Berlin. The programme combined live-streamed lightning talks with structured in-person breakout work, bringing together technical architects, developers, metadata

specialists, and infrastructure representatives from NFDI consortia and Base4NFDI services. The full programme and speaker slides are available at <https://events.hifis.net/event/2908>.

## Day 1: Interoperability in Practice

The first day opened with lightning talks covering RSpace (ELN capabilities and workflow design), ORCID (identifier infrastructure), NOMAD (metadata capture for computational research), and TS4NFDI (harmonised terminology access and mapping services). These sessions grounded the workshop in live implementations and set the stage for the afternoon breakout sessions.

Two breakout streams were organised around the key workflow types.

### ELN Stream

The ELN breakout session conducted a practical deep dive into the technical pathways of research data documentation. Using RSpace as a reference ELN, the ELN breakout session ran a three-step methodology:

1. **Metadata Review:** Participants conducted a comprehensive review of DataCite fields for datasets and instruments, evaluating whether RSpace currently captured each field and through which technical mechanism (manual entry, automated device output, or external import).
2. **Obligation Classification:** For each field, participants classified metadata as **Mandatory**, **Recommended**, or **Optional**, while flagging areas where current ELN metadata was insufficient for high-quality PID registration that would support provenance, data lineage, reuse, and impact analysis.
3. **Data Lineage Analysis:** Participants mapped the origin and flow of each metadata element, distinguishing between manual entry, automated device output, and external imports, then traced its downstream path to PID services, repositories, and APIs. By visualizing these pathways with flow diagrams, the group identified critical gaps and disrupted workflows where metadata is currently lost or requires manual intervention.

The exercise surfaced a finding that became a central theme of the entire workshop: the primary issue is not missing fields but missing structure. **Creator** names were recorded without ORCID links. **Affiliations** appeared without ROR identifiers. **Subject** terms were present without controlled vocabulary URIs.



Metadata that looks complete at the point of entry loses interoperability at every system boundary, not because information is absent, but because it lacks the structure required for machine-actionable transfer. This framing clarified the distinction between metadata presence and metadata quality, and pointed directly to the solution: structured, identifier-based capture at source. In this context, the integration of external PID services and standardised vocabularies emerged as a key opportunity to automate enrichment, reduce ambiguity, and improve the reliability of metadata across system boundaries.

### *DMP Stream*

The DMP breakout session focused on the structural alignment between planning tools and global metadata standards. Using the RDMO framework as a baseline for examining alignment with the DataCite schema and the RDA Common Standard for maDMPs<sup>7</sup>. A central task was resolving ambiguities in how core entities are defined, particularly the distinction between a “dataset” and an overall “project” as metadata objects. The session confirmed that planning tools contain rich contextual information that is not structured for automated transfer or persistent linking: funder details exist but without ROR identifiers, contributor records exist but without ORCIDs, subject classifications exist but without vocabulary URIs. The discussion also underscored the importance of lifecycle-aware metadata, particularly versioning and provenance, for plans that evolve continuously rather than representing fixed descriptions.

### **Day 2: Standards, Quality, and Implementation**

The second day shifted focus toward standards consolidation, metadata quality metrics, and practical implementation planning. The morning began with strategic discussions on developing **training materials**, followed by a second round of lightning talks covering specialized topics such as PIDs for instruments (PIDInst), the DMP4NFDI base service, and the MaLDReTH Map as a reference for tool interoperability. Subsequent sessions involved deep-dive technical discussions centered on PID metadata completeness and the practicalities of software integration. These sessions allowed participants to bridge the gap between **high-level requirements and concrete implementation steps**.

The final phase of the workshop involved a deep-dive "scorecard" exercise where participants chose a research object, an ELN entry (Collection), a DMP record, or a

---

<sup>7</sup> Miksa, T., Walk, P., & Neish, P. (2020). RDA DMP Common Standard for Machine-actionable Data Management Plans. Zenodo. <https://doi.org/10.15497/rda00039>



custom research entity, and worked through a set of KPI cards. For each card they discussed relevance, mapped the relevant DataCite 4.6 property paths, noted PID/standardisation options, and assigned an importance score: 20% = nice-to-have · 50% = helpful · 80% = very important · 100% = critical. Groups then agreed on top-3 priorities, one blocker, and one quick win:

- **The ELN Group** identified **re-use licenses, provenance** (metadata origin), and **persistence** as their top priorities.
- **The DMP Group** prioritized **funding information** and **versioning** as critical for tracking data provenance.

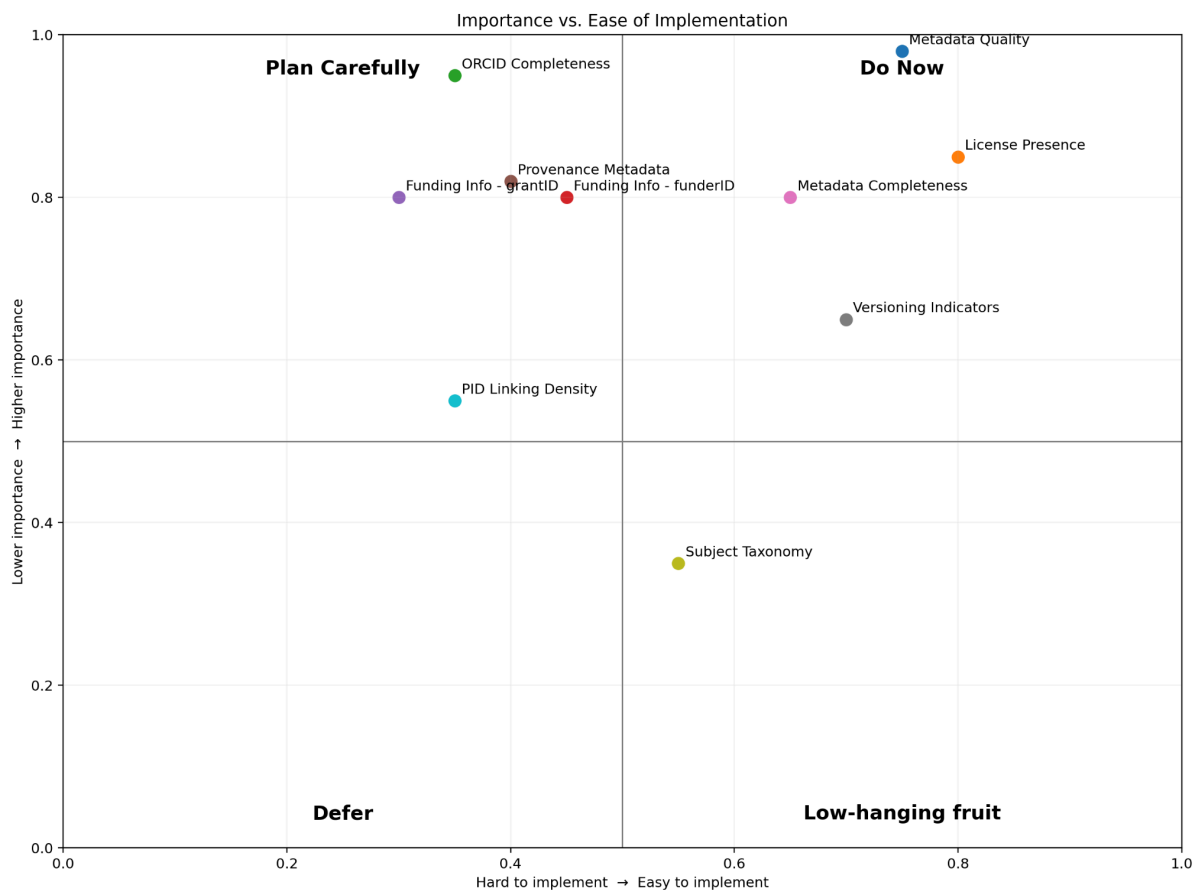
**The Workshop Group** discussed the nuances of project-level PIDs like **RAiD**, and specific information requirements of projects as a resource type, e.g. how to capture start/end dates and how to handle metadata updates for projects without third-party funding. The table below synthesises the importance scores across all three groups.

KPI	DMP	Projects	ELN	Consensus
Metadata Quality	100%	100%	100%	Universal
ORCID Completeness	100%	100%	80%	Universal
License Presence	100%*	100%	80%	Universal (*context)
Metadata Completeness	80%	80%	80%	Universal
Funding Information	80%	100%	80%	High
Provenance Metadata	100%	50%	50%	Object-dependent
Versioning Indicators	100%	n/a	varied	Object-dependent
Subject Taxonomy	20%	50%	20%	Lower priority

**Table 1.** Importance scores assigned by the DMP, Projects, and ELN groups for each KPI dimension. \*License presence was rated 100% in open/reuse contexts and near 0% for purely internal archival use; the score shown reflects the open-context rating.

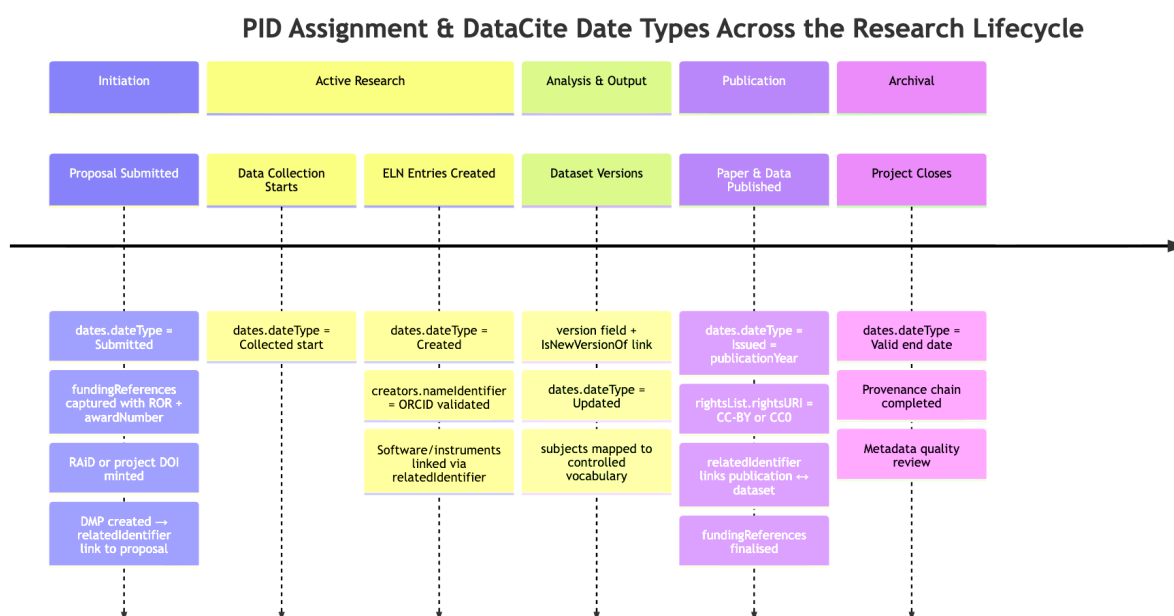
Three findings cut across all groups. **First**, object type determines what matters: the same DataCite property can be critical for a DMP and effectively irrelevant for a project record. A single universal metric framework will not work, per-object-type **profiles are needed**. **Second**, ORCID is universally critical but hard to capture: all groups rated creator identification at 80–100%, but the shared blocker was collecting ORCID iDs at the point of record creation rather than retroactively. **Third**,

funding structure requires two PIDs: the workshop converged on needing both a funder identifier (ROR or Crossref Funder Registry) and a grant or award number (captured together via RAiD/DOI where applicable), capturing only one was repeatedly flagged as insufficient.



**Figure 3. Importance versus ease of implementation:** where to focus integration effort first.

One question that crossed all groups concerned how data maps to project time: what does “publication year” mean for a DMP? When should a project-level identifier be assigned? These questions point to a gap in how DataCite date types are currently applied across the research lifecycle, and directly inform the scope of the incubator programme.



**Figure 4.** PID assignment and DataCite date types across the research lifecycle.

The workshop reached a clear shared conclusion: further discussion was not the limiting factor. The community had converged on the technical priorities, the correct reference schemas, and the right role for PIDs in DMP and ELN workflows. What was needed was a practical implementation step, one that would demonstrate, in a working system, that early-capture metadata with embedded controlled vocabularies can survive system boundaries and reach repositories with lineage intact. That conclusion directly motivated the **incubator programme** described in the following section.

## 6. The Next Phase: Incubator Projects with TS4NFDI

### 6.1 From Community Alignment to Technical Prototyping

Building on the requirements identified across the focus group meetings and the Berlin workshop, PID4NFDI and TS4NFDI are jointly developing two complementary incubator projects. The programme represents the transition from community alignment, understanding what needs to be built and why, to practical prototyping: demonstrating that it can be built, and leaving behind reusable infrastructure for those who follow.

The collaboration rests on a shared recognition that persistent identifiers alone are not sufficient for sustainable interoperability. Tools also require centrally maintained, machine-actionable vocabularies and schema crosswalk that preserve meaning as metadata crosses system boundaries. The DataCite Metadata Schema 4.6 serves as the canonical reference model, not to enforce uniformity, but to provide a stable foundation against which domain-specific schemas can be aligned. TS4NFDI contributes three core infrastructure components: the **Cocoda** Mapping Service for creating and maintaining [JSKOS](#)-format schema crosswalk; the [Terminology Service Suite](#) (TSS) for embeddable vocabulary lookup widgets; and the API Gateway for federated querying across distributed terminology services.

## 6.2 Incubator Project 1: DataCite Terminologies in TS4NFDI

The first project operationalises the DataCite metadata model and its controlled vocabularies within the TS4NFDI infrastructure. The objective is to make these vocabularies available as centrally managed, machine-actionable resources accessible through standard APIs and reusable interface components.

Key activities include curating the core DataCite controlled vocabularies, [relationType](#), [resourceTypeGeneral](#), [contributorType](#), [dateType](#), and [relatedIdentifierType](#), in interoperable SKOS/JSKOS format with stable identifiers, human-readable labels, and formal definitions. In parallel, curated crosswalk mappings between DataCite and related schemas (Schema.org, DCAT, and PROV-O) are developed and maintained via the Cocoda service, enabling semantic alignment without requiring system-specific implementations. The vocabularies are exposed through the TS4NFDI API Gateway and integrated into TSS widgets, allowing any NFDI tool to incorporate standardised terminology lookup directly in its interface. Governance structures defining long-term maintenance, versioning, and coordination with international initiatives are established as part of the project.

The expected outcome is a centrally managed terminology and mapping layer that enables consistent, machine-actionable use of DataCite semantics across NFDI systems, applicable to any tool that implements the integration, not only RSpace.

## 6.3 Incubator Project 2: Semantic ELN Integration (RSpace)

The second project integrates DataCite metadata properties and controlled vocabularies into RSpace ELN workflows, demonstrating how a production ELN can

capture standards-aligned, PID-ready metadata with minimal overhead for researchers.

Rather than modifying RSpace's underlying data model, the approach embeds TSS widgets directly into existing metadata fields, adding semantic structure at the user interface level. Researchers select standardised terms with stable URIs via the TS4NFDI API Gateway instead of entering free text. The integration focuses on a curated set of high-impact fields: relation types (`relationType`, enabling values such as `IsDerivedFrom` and `IsSupplementTo`), resource types (`resourceTypeGeneral`), and structured capture of ORCID and ROR identifiers.

A key demonstration use case is data lineage. When a researcher selects `IsDerivedFrom` through a TSS widget to record that an ELN entry derives from another dataset, the resulting metadata is consistently interpretable across DataCite, Schema.org, and PROV-O, three frameworks that otherwise express equivalent relationships differently. End-to-end, the workflow runs from ELN record creation with enriched metadata through export to a repository with preserved, machine-readable lineage.

The expected outcome is a production-oriented prototype demonstrating how ELNs can support standards-aligned, PID-ready metadata with minimal user overhead, a practical blueprint for adoption by other NFDI tools.

## 6.4 Timeline, Sustainability, and what comes Next

The PID4NFDI–TS4NFDI incubator project runs from April to October 2026, within the active window of PID4NFDI's current funding phase. This timeline is a firm constraint: core personnel contracts conclude in October 2026 and project funding closes in December 2026, making this period the final opportunity to consolidate the community's analytical work into maintained infrastructure. Vocabulary curation and first-pass mappings are targeted for Q2 2026, with testing, and documentation following in Q3 to Q4.

Running in parallel, the collaborative groundwork between PID4NFDI and RSpace, identifying and mapping the relevant metadata fields, is already underway. From Q3 2026 the RSpace–TS4NFDI incubator project is expected to commence, at which point widget integration and end-to-end workflow demonstration become the focus, building directly on the mapping work done in the preceding months.

The DMP strand, deferred pending the mid-2026 maDMP standard update and DMP4NFDI capacity, is expected to reconnect with these infrastructure outputs once conditions allow. The vocabulary and crosswalk layer built in Incubator Project 1 is intentionally designed to serve both ELN and DMP integration equally: the same machine-actionable DataCite terminologies that RSpace will consume are directly applicable to future RDMO plugin development. The incubator work is not a detour from the DMP goals established in the focus groups, it is the shared foundation on which both strands will eventually converge.

The results from both projects will be documented as reusable blueprints: published mappings in the Cocoda service, and guidance for other NFDI tools seeking to align with DataCite semantics. The focus groups established through this programme remain the primary feedback mechanism for iterative refinement. The long-term aim is a research data infrastructure where DMPs, ELNs, and downstream repositories share a common semantic layer, one where high-quality metadata, captured once at the point of creation, flows without loss or manual re-entry across the full research lifecycle.

## References

El-Gebali, S., & Böhm, J. (2025). *Landscape Analysis of PID Practices in NFDI* (1.2). Zenodo. <https://doi.org/10.5281/zenodo.15689799>

IBM. (n.d.). *What is data lineage?* Retrieved April 16, 2026, from <https://www.ibm.com/think/topics/data-lineage>

Miksa, T., Walk, P., & Neish, P. (2020). *RDA DMP Common Standard for Machine-actionable Data Management Plans*. Zenodo. <https://doi.org/10.15497/rda00039>

PID4NFDI. (2025). *Data Collection & Mapping Baseline* [Unpublished internal spreadsheet]. <https://docs.google.com/spreadsheets/d/1YS0ZC7dCFxN3DLzmR-qlyye5U1AT-k2ID2xidGxEEr8/>

PID4NFDI. (2025). *Background reading Mappings, Crosswalks, Ontologies* [Unpublished internal document]. <https://pid.services.base4nfdi.de/blog/crosswalks/>