Case Statement: PID Kernel Information profile management WG

(aka PID Kernel Information WG #2)

Co-chairs: Tobias Weigel, Beth Plale, Jens Klump

Motivation and scope

The PID Kernel Information WG produced a recommendation that contains i) guiding principles for identifying information appropriate as PID Kernel Information, ii) an exemplar PID Kernel Information profile, iii) several use cases, and iv) architectural considerations. PID Kernel Information is defined as the set of attributes stored within a PID record. It supports smart programmatic decisions that can be accomplished through inspection of the PID record alone.

At the RDA P13 PID Kernel Information WG session, attendees expressed enthusiasm over the PID KI work. While the WG feels that the exemplar profile reflects a consensus decision while at the same time satisfying the guiding principles, it does not preclude other profiles. But a global technology, such as the PID KI intends to be, that is without considered organizational governance or management will never be adopted except in highly leading edge (aka research) settings.

Thus the PID KI WG agreed at the conclusion of the P13 session that there was need to examine the governance and management of some small number of globally relevant PID KI profiles. Without further guidance on how to manage profiles, there is the risk of wide-scale proliferation of overlapping or incompatible profiles, which could significantly hamper long-term realization of coherent middleware that supports Kernel Information. As the PID KI depends on a type profile to be interpretable, it requires the existence of a Data Type Registry to store the type definition of the PID KI. In fact, both PID KI and DTR are part of the same digital object ecosystem. Thus the issue of governance must be coordinated with the DTR#3 group. A follow-on WG additionally presents an opportunity to further define the boundary conditions for such middleware and foster alignment across disciplines and regions.

The WG lives in the context of

- Multiple use cases from disciplinary adoption and project work
- Relevant work inside RDA (other WG/IG) and outside RDA (W3C PROV, SemWeb/LD in general)

Objectives

The WG will work towards the following objectives:

1. Life cycle model defined for KI profiles and mechanisms (principles, processes, tools) through which KI profiles can be defined and governed. Define profile metadata and how to encode a profile.
2. Baseline KI architecture extended to match the needs of profile management, for example, by including profile registries and connectors.
3. Describe the technical interface for interaction with profile registries, most likely based on the DTR WG recommendation and API. Preferably, a separate new API does not need to be defined.
4. As the PID KI depends on a type profile to be interpretable, and both PID KI and DTR exist in the same digital object ecosystem, the governance and management of PID KI profiles should be synergistic with the governance and management of Data Type Registries. Objective is coordinated guidance on governance between this WG and the DTR#3 group.
5. Facilitate coordination with other RDA groups mentioned further below more generally.

Value proposition

Cyberinfrastructure providers will benefit from the architectural reference model, which makes collaboration for development and use of shared infrastructure components using PID KI profiles easier. The governance mechanisms provided by the WG are a necessary element for serving KI profiles in operational settings.

Tools and services builders will benefit from the availability of well-defined, well-managed and interoperable KI profiles as they can rely on agreed profiles to underpin specific tasks in data workflows. A common approach for interfacing with profile registries can reduce development costs, for example by sharing software library development for registry clients.

As a result of the combined efforts by cyberinfrastructure providers and service/tool builders, **scientific users** will benefit from the better availability of information across data life cycle stages through the connections made by the PID KI graph. The adoption of profile management can make the links in the graph more coherent.

Data producers, particularly if organized in research infrastructures, could benefit from having a wide range of KI profiles available, with ensured interoperability between them and related services. Potential needs and usages will differ between research disciplines, but knowing which profiles are supported by which repositories could help in the process of identifying the optimal storage and cataloguing facility for a given dataset.

Stakeholders and adoption

The topic is of relevance to multiple stakeholders, which will be sought out. Initially, these are:

- **The NSF-funded eRPID project:** The predecessor RPID project evaluated and prototyped usage of the early Kernel Information recommendation and its findings within the project context. The eRPID successor project will continue this and provide two-way interaction on practical use cases and needs, and be informed by the PID KI WG outcomes.
- **EUDAT:** The operational B2HANDLE service is relying on Kernel Information as an essential element to facilitate cross-service integration at e-infrastructure level, largely hidden from users as is in line with the KI vision. The latest integration activities (e.g. B2SHARE, OneData) introduce a complexity level where profile governance would be highly beneficial to support long-term stability.
- Similar to EUDAT, **ePIC/GWDG** take further interest in the KI concept to underpin identifier-level metadata management with a stable conceptual framework, and participate in the discussion of governance processes with the eyes of a potential adopter of these processes at the ePIC management level.
- **International GeoSample Number (IGSN):** IGSN interest in the KI concepts was renewed at and after the P13 session. IGSN may benefit from KI to provide better integration with e-infrastructure services, and a discussion on profile governance may be a key requirement for adoption actions.
- **DiSSCo:** As a new research infrastructure in Europe connecting natural scientific collections of more than hundred collecting holding institutes across Europe, DiSSCo has a potential interest in the KI concept to inform its future e-infrastructure strategy. DiSSco is in conversation with IGSN to develop a joint Handle-based approach for specimen object identifiers, and this has the potential to lead to billions of new Handles. The KI concepts and governance discussions are relevant and crucial for long-term stable e-infrastructure services in this context.
- **Deutsches Klimarechenzentrum (DKRZ):** DKRZ has introduced PID management services into the wider IS-ENES/ESGF CMIP6 climate data infrastructure, which are now based on PID Kernel Information and other related RDA outcomes. Notably, the recommendations of the Research Data Collections WG have also been put to practice, and the resulting solution may also inform KI discussions. With the first generation of these services operational, first practical experience indicates that integration beyond the e-infrastructure and/or community boundary requires solid KI profiles, and such may easily fail if governance were not addressed. DKRZ follows these discussions with a specific view on the future IS-ENES infrastructure and potential next generation of services for CMIP7.

To facilitate adoption in practice, the group's work needs to be informed by specific exemplary cases from these stakeholders, ensuring that both the group's work is grounded in practical reality and that the adoption barrier is lowered, turning stakeholders into 'guinea pigs'. The following examples illustrate where the availability of profiles and their management have or can have benefits in actual usage:

- **eRPID example:** The RPID demonstrator uses the PID KI exemplar profile already registered in DTR. This will be further extended as part of eRPID. The findings of RPID/eRPID in a prototypical service environment can inform the PID KI discussions for matching practical use cases and infrastructure needs.
- **EOSC example:** The constituent technical services of the European Open Science Cloud (EOSC), notably from EUDAT and EGI, could benefit from better structure of PID records through profiles and the organizational streamlining that adherence to profiles fosters. Profiles can be specific and characteristic for different life cycle phases (preliminary data sharing, data archival). In particular, the ECAS data processing services is prototyping basic data lineage tracking through a basic data input-processing-output workflow. The processing environment needs to become aware of an object's context (e.g., are they shared preliminary data or archived data?) to a) inform the user of any underlying implications for their data analysis; b) record provenance with confidence.
- **Service orchestration example:** This is a use case discussed in past iterations of the DTR WG with wider general applicability in service-oriented architectures compliant with the Data Fabric model. If data objects and service objects receive PIDs and PID KI, then it may be required to filter out services from among all objects in order to let an orchestrator determine precisely which objects can be put to which services and how service chains can be generated dynamically. Profiles are an essential ingredient to make the filtering work.

Relation to other RDA efforts

As reflected in objective #4, the group's work should be closely coordinated with DTR #3 as this group deals with a specific application case for DTR, and the PID KI and DTR both exist within the same digital object data ecosystem (one that builds upon, but is not limited to, the Handle system for PID resolution).

In addition, the use cases and stakeholder perspectives must be further taken in by connecting to the PIDs for Instruments WG, the emerging group on Interoperability of Observable Property Descriptions, the Biodiversity Data Integration IG and the PID IG. The latter is of specific importance as a forum to engage new adopters, both from use and added-value perspectives as well as infrastructure and service provisioning perspective. In this context, discussions between FREYA and RDA as part of the PID IG group are relevant to the Kernel Information context, and will be actively followed on by the group.

Within the Instrument PID group scope, metadata under consideration is primarily based on DataCite kernel, which is much more extensive and geared towards different usage scenarios than PID KI. However, there is also ongoing discussion about implementing a more “generic Handle-based registry” profile implementation. As such, they may form a candidate for a profile for getting instruments and their products in the PID graph and enable filtering per instrument/measurement campaign/sample ‘type’ (physical, virtual, or their subtypes). A discussion along these lines will be taken up during early WG lifetime.

Work plan and milestones

The WG will first assemble requirements for profiles and supporting services. Then, it will work towards the objectives in parallel.

Assuming a first formal session at P14, the group work aligns with the following milestones:

- P14: Presentation of group scope and goals to RDA community, engagement with additional use cases, discussion of requirements for profile management.
- P15: Analysis of use cases and requirements complete and first draft of boundary conditions for profile management presented. Intake of community input to profile management and governance principles, review of possible frameworks.
- P16: Presentation of first draft of full profile management framework and readily adoptable profiles. Discussion of implementation gap analysis.
- P17: Delivery of outputs.

In the end, the WG will deliver a recommendation for KI profile management, example profiles defined together with potential adopters, profile registry interface specification (if needed).

Initial supporters

Tobias Weigel, DKRZ
Beth Plale, IU
Larry Lannom, CNRI
Ulrich Schwardmann, GWDG
Jens Klump, CSIRO
Mark Parsons, RPI
Maggie Hellström, Lund University