PID Kernel Information WG

Co-chairs: Beth Plale (IU – Data to Insight Center), Tobias Weigel (DKRZ)

As has been discussed in several awareness raising sessions with the RDA community at RDA Plenaries 7 and 8 in special sessions carried out in the Data Fabric IG [2, 3] and at two extracurricular RDA events [4, 5, 6], there is a gap to realization of the power of the combined RDA Recommendations (PID Information Types Recommendation and Data Type Registry Recommendation). This gap, described in more detail below, is in coming to consensus around what makes up PID Kernel Information. Convening interested parties in the RDA community to define PID Kernel Information, and having a PID Kernel Information Recommendation in the RDA recommendation suite will allow adopters of the core RDA recommendations (PIT Data Types [10] and Data Type Registry [9]) to take full advantage of the unique power that these recommendations have when combined.

We propose a short-lived working group, whose work is as proposed below, to converge through a series of discussions and a narrowing process on the smallest number possible of versions (or profiles) of PID Kernel Information (one is ideal but not likely). *The PID kernel information versions/profiles, which will be the primary output of the WG, will be focused on the needs of finding and using research data. That is, seeking applicability of the profiles outside the R&D sector is out of scope of this particular working group activity.* We anticipate members of the group offering representations (i.e., implementations) as well. This working group is a formalization of an already underway effort initiated as subgroups of the Data Fabric Interest Group at Plenary 8.

PID Kernel Information

PID Kernel Information is a small amount of information about a data object to which the PID refers. This associated information resides within the PID resolution system itself. The information is domain agnostic as to do otherwise results in splintering into numerous profiles which would negate the benefit of common information. A close analogy to PID Kernel Information is the header information in a TCP packet: the information is minimal, domain agnostic, and critical for decision-making. It resides within the network (aka, is packaged as part of every TCP packet.)

The DOI[1] system collects metadata about an object when the object is registered, and has a minimal metadata set that it requires to facilitate recognition and interoperability (https://www.doi.org/doi_handbook/4_Data_Model.html#4.3.1). This functioning system has shown to be highly useful in publishing circles. As has emerged repeatedly in RDA, the DOI system is (currently) not considered the first choice to address the dynamic and granular needs of active research and the proliferation of observational data from instruments, devices, and sensor networks. This has reasons related to implementation, legacy and the focus of the specific user community, though it is not a quality necessarily inherent in the DOI system's conceptual design. But independent from possible future activities and changes within the DOI system and its user community, the momentum for Kernel Information profiles needs to be brought in by RDA at this point to possibly induce benefit also for DOI stakeholders.

[1] Note that DOIs are handles from a specific well-organized community with a common approach and a strong social contract and that by the term 'handle' we mean those handles that are not also DOIs, e.g., those used by DSpace or EUDAT.

The DOI community, however, was first to recognize the value of minimal metadata associated with a DOI. According to people familiar with the effort, ‘associated’ is stretched to mean somehow reachable, e.g., appearing on the default or ‘landing’ page pointed at by the DOI. This effort within the DOI community, however, has always been more aspirational than operational. We share in common a reference to this minimal metadata as “*kernel information*” with our terminology being “PID kernel information”.

The benefit of PID kernel information has been articulated in such activities as Haga [7] and Weigel [11] and is not repeated here save to say that a motivating use case for the effort is this:

> Example Use Case: Suppose we live in a world in the not too distant future where an internet-scale service is handed a list of 100,000,000 PIDs that could describe digital objects from the entertainment industry, from Internet of Things, research or physical objects. How does the service quickly sort through huge lists of PIDs to find the research data? How can that be done without making 100,000,000 calls to repositories? And further, suppose it can not only pick out the research data, but then make a simple determination of whether it can trust and designation. This is not possible today, but the work of this WG will take us one step closer.

More specifically, machine uses of PID services rely on standardized and domain agnostic results of PID resolution. We posit that resolution results should include more information than just the PID. This does not mean that all resolution must return exactly the same kind of information, which would be too crude an approach to cover the varieties of PIDs and their uses, but that standardized type information can be associated with PID records and included in resolution results. These types we are assuming to be registered in a Data Type Registries (DTR). As with DNS, PID resolution must be extremely fast and reliable, hence additions to PID information must be balanced against the performance tradeoffs. When meaningful decisions can be made about a data object simply looking at nothing more than the kernel information of a PID, a new economy of services can grow.

The objective of this WG is to assess multiple profiles gathered through these activities and improve on them, compile a core profile from them and work towards practical adoption. The group will use a provenance flagship use case to demonstrate cross-community usage based on minimal information contained within the core profile. The core information can possibly be enriched to full W3C PROV compatible provenance that is interpreted by higher provenance enablement layers.

Milestones, Objectives and Deliverables

At RDA Plenary 8, our group formed 4 subgroups, each with a different perspective to identify suitable PID Kernel Information. These groups are:

Science data: Consumer perspective	Science data: Provider perspective
Humanities data: Consumer perspective	Humanities data: Provider perspective

Table 1: PID Kernel Subgroups from Plenary 8

Milestones

At the Plenary 8 meeting we had at least 3 individuals sign up in each of the 4 groups, making the breakdown viable to continue with.   Through Winter '16, we have held monthly phone calls with the goal to arrive, within each subgroup, with a single definition of suitable PID Kernel Information. The milestones for the effort can be captured in Table 2.

Fall'16 Plenary 8	Sign up at least 3 individuals in each subgroup.	Completed Fall '16
Winter '16	Monthly phone calls for each subgroup. Arrive at single profile for PID Kernel Information per subgroup.	Ongoing
April '17 Plenary 9	Bring 4 PID Kernel Information profiles together in a session at RDA to discuss. Determine an approach to harmonization.	TBD
Summer '17	Work through issues of harmonization; work through issues of representation (implementation) particularly in Data Type Registry; address provenance enablement. Provide a perspective for integration with Linked Data approaches.	TBD
Fall'17 Plenary 10	Bring draft final conclusions to work group for discussion.	TBD
Dec'17	Publish product of effort for RDA feedback	TBD
Mar'18 Plenary 11	Incorporate feedback from community, publish results and close down group	TBD

Table 2: PID Kernel Information group Milestones

Objectives

The objectives of the group are:

1. Work out the terminology and data model for PID driven data discovery. This is scoped very closely to the goal at hand, but needed for assumptions about what goes into PID Kernel Information
2. Resolve to small number of PID Kernel Information instances
3. Work through implementation/representation issues as a way to test the goodness of the instances selected in step 2.
4. Work through use cases offered by group participants as a way to further test the goodness of the instances selected in step 2
5. Write up and present results in forms accessible to the RDA community to seek broadest possible input

6. Offer a forum for discussing and developing additional community profiles and use cases

The group aims to finish its work by March 2018.

The group relies on its members to bring to its attention existing approaches, such as DOI's kernel information efforts, and will remain aware of these efforts so that our effort can be set properly in the context of other work. For instance, a use case that has emerged is universal provenance. For this use case, the group will particularly ensure compatibility with the PROV recommendation to support interoperability. As the information gathered in the PID profiles is likely to be insufficient for a full PROV representation, the group should define ways to integrate richer sources, for example by describing an architecture where higher layers provide such information, with the underlying PID record layer as a fallback and a method to efficiently address large numbers of objects.

Deliverables

D1. Terminology and data model for PID driven data discovery
D2. Conceptual description of group agreed upon PID Kernel Information instances
D3. Actionable form of PID Kernel Information instances (scripts, schemas)
D4. Report and presentation material to enable community assessment and feedback and for archival purposes

The group intends to work through dedicated sessions at each plenary and regular virtual meetings between plenaries. The final output of the group should include documentation on intermediate results, most importantly the individual profiles gathered and discussed at its meetings, to enable follow-up groups to understand the whole process and potentially extend it in directions not foreseen during the group's lifetime.

Value proposition and adoption

Data discovery ecosystem. The biggest beneficiaries of PID Kernel Information are data discovery clients that operate at internet-scale speeds to act on large lists of PIDs. Today it makes no sense to transact in lists of PIDs (acquired such as through a deep web harvest) because services that can act intelligently on these lists cannot be built. PID Kernel Information addresses this shortcoming through small amounts of information embedded with the PID, and because it is embedded, it is available to internet-speed services that route and filter based on content. The services can, for instance, separate research data from non-research data, and within the PIDs for research data, can further make simple determinations of trust using the provenance embedded in the record, thereby enabling a service to hand to another service a list that the creating service has reasonably good confidence in.

Client tool builders (e.g. discovery, content routing, search, provenance) will benefit from internet-scale efficiency in interfaces and protocols established to make decisions on PID Kernel Information. This kind of support, which is completely absent from a data ecosystem except in pockets, will enable a whole new suite of client services to emerge. One particular application may be to rely on Kernel Information to let an agent configure and conduct data processing jobs in an automated way requiring minimal user interaction, which is a general scenario put forward by the Data Fabric IG under the concept of Type-Triggered Automated Processing (T-TAP).

Scientific users will benefit from widespread availability of PID Kernel Information in that it will allow them to make determinations of trust and use of existing digital data in ways that they are not able to today.   For instance, a provenance use case that we are working with gives minimal provenance information that is uniformly available and widespread rather than in-depth information available only at few places.

The working group will conduct outreach as means allows in support of getting word out. It will develop archive oriented materials for consumption in places where travel means do not exist.

Interaction with other groups

This WG has already or foresees useful interactions with the following RDA groups:

- Data Fabric IG: feedback on the practical experience made with the data fabric component concept and obtain future directions about its refinement
- Provenance IG: attract and inform possible adopters and gather input for the main use case
- Data Type Registries WG: keep building towards a profile registration and discovery mechanism
- Metadata IG: while the Kernel PID Information is highly specialized, it is still a form of metadata, and to the extent it can align with categorization and terminology defined, we should.
- Collections WG: there are overlaps here in representation of collections
- Brokering WG: a broker is a server that maintains minimal information about objects and as such has overlaps with PID kernel information that we will explore

The group also seeks interaction outside RDA within the limited scope possible. Of particular concern are the aforementioned efforts in the DOI community that aim for consideration of kernel information profiles to address the “zero knowledge” object problem, following discussions at the Pidapalooza event in Reykavík in November 2016. WG members Jonathan Clark and Larry Lannom are involved in this effort.

There is currently no direct interaction with members or potential adopters from industry; however, given the activities at the DONA level regarding IoT applications and the possible industry involvement at P11, the WG has a perspective on reaching out to such potential members, though this largely depends on how far the early ideas of profiles can actually be taken towards such concrete applications in the limited timeframe of the WG.

Membership

Bridget Almas, Perseids Project, Tufts University, USA
Beth Plale, HathiTrust Consortium and Data To Insight Center, Indiana University, USA
Alex Thompson, iDigBio, University of Florida, USA
Tobias Weigel, German Climate Computing Center (DKRZ), Germany
Jim Duncan, Vermont Monitoring Cooperative, USA
Stuart Chalk, University of North Florida, USA
Kei Kurakawa, National Institute of Informatics, Japan
Ulrich Schwardmann, GWDG, Germany

Adopters

The below list of projects have agreed to engage in adoption of the primary outcomes of the WG (see D2 and D3 of Deliverables).

- Beth Plale, HathiTrust Consortium and Data To Insight Center, Indiana University, USA. Adopt into IU SEAD Cloud
- Alex Thompson, iDigBio, University of Florida, USA. Adopt into iDigBio.
- Tobias Weigel, German Climate Computing Center (DKRZ), Germany. Adoption into DKRZ/ENES and EUDAT
- Ulrich Schwardmann, GWDG, Göttingen. Adoption as MPA/ePIC.
- Bridget Almas, Perseids Project, Tufts University. Adopt into Perseids Platform

Adoption plan details:

As we have discussed with the project representatives, we acknowledge "adoption" as a process, drawing from a 1981 study on diffusion by Beal and Bohlen (1981) who observe that "the process by which people accept new ideas is not a unit act, but rather a series of complex unit acts." Acceptance of a policy, practice, or technology is thus a process having distinct stages through which a person passes: awareness, interest, evaluation, and finally integration into a production environment. Adopters have agreed to undertaking evaluation.

Adopters start their evaluation process latest at month 12 when the final group deliverables are available. Evaluation means that adopters will set the group deliverables against the particular requirements of their practical environment, existing or future applications and other considerations relevant to their context. The evaluation process is by definition open ended and adopters can drop out at any point in time. However, adopters are encouraged to report back to the group and RDA as a whole as part of the post-WG phase, and may contribute to a status report envisioned to take place at an RDA plenary about 12 or 18 months after the group has delivered. This status report should cover evaluation results, actions taken and planned for the future.

References

1. I. Suriarachchi and B. Plale, Crossing Analytics Systems: A Case for Integrated Provenance in Data Lakes, *IEEE 12th Int'l Conference on e-Science*, Baltimore, Oct 2016
2. L. Lannom, P, W.ttenburg, B. Plale, IG Data Fabric - Data Fabric and Common Components - state and perspectives, RDA Plenary 7 Tokyo Japan, 1-3 Mar 2016, Breakout Session 7
3. B. Plale, T. Weigel, L. Lannom, PID centric fabric constructed piece by piece, RDA Plenary 8 Denver CO, 15-17 Sep 2016, Breakout Session 2
4. T. Weigel and B. Plale, Building a concrete configuration: viewed through the lense of direct and indirect approaches, RDA PID Training, Garching, Germany, Aug 2016
5. B. Plale, Role of PIDs in Data Quality: a Tutorial, Data Quality in an Era of Big Data, Bloomington, IN Sep 2016
6. B. Plale, Power of PID Kernel Information, RDA WG meeting, NIST, Maryland, Dec 2016
7. B. Plale, Jason Haga, Inna Kouper, Bringing visibility to food security data results: Harvests of PRAGMA and RDA, RDA Plenary 8 RDA Recommendations and Adoption Plenary Session, Denver CO, 15-17 Sep 2016
8. Transmission Control Protocol, https://en.wikipedia.org/wiki/Transmission_Control_Protocol
9. L. Lannom, D. Broeder, G. Manepalli, L. Bartolo, C. Blanchi, J. Braswell, W. Chang, S. Cox, T. DiLauro, J. Erikson, A. Fillinger, P. Fox, B. Hadden, M. B. Jones, X. Ma, N. Paskin, A. Powell, S. Richard, U. Schwardmann, J. H. Scott, Z. Trautt, R. Tupelo-Schneck, T. Weigel, P. West, S.

Youssef, T. Zastrow, S. Zednik, Data Type Registries Working Group Output, Research Data Alliance, Aug 2015, http://dx.doi.org/10.15497/A5BCD108-ECC4-41BE-91A7-20112FF77458, https://rd-alliance.org/group/data-type-registries-wg/outcomes/data-type-registries

10. 10. Weigel, T., T. DiLauro, T. Zastrow, PID Information Types WG final deliverable. Research Data Alliance, Aug. 2015. dx.doi.org/10.15497/FDAA09D5-5ED0-403D-B97A-2675E1EBE786, https://rd-alliance.org/group/pid-information-types-wg/outcomes/pid-information-types
11. 11. Weigel, T. Persistent Identifiers for Earth Science Data Management. Dissertation, Universität Hamburg, 2016. urn:nbn:de:gbv:18-78448. http://ediss.sub.uni-hamburg.de/volltexte/2016/7844/