What is the Problem?
Numerous systems and providers to register and resolve Persistent Identifiers (PIDs) for Digital Objects and other entities have been designed in the past and are used today. However, almost all of them differ in the way they allow researchers to associate additional information, such as for proving identity and integrity with the PID. For application developers this is an unacceptable situation, since for all providers a different Application Programming Interface (API) needs to be developed and maintained. Given that a researcher has found a useful file, but first wants to prove whether it is indeed the same stream of bits after some years, he should be able to request the checksum independent of the provider holding the PID. How should he do this not knowing whether the provider offers this information and if so, how to request it? We can overcome such extreme inefficiencies only if all providers agree on a common API, register their information types in a common data type registry and agree on some core types, such as the checksum.

What were the goals?
The goals of this WG were:
- Coming to a core set of information types and register (and define) them in a commonly accessible Data Type Registry
- Providing a common API and prototypical implementation to access PID records that employ registered types

What is the solution?
The PIT group accomplished the following:
- Defined and registered a number of core PID information types (such as checksum)
- Developed a model to structure these information types
- Provided an API, including a prototypical server implementation that offers services to request certain types associated with PID records by making use of registered types.

Due to high demand, a variety of trusted PID service providers have been set up already, yet all of the different attributes associated with the registered PIDs make life of a software developer a nightmare. We need to harmonize the major information types and suggest a common API, so that if we request the checksum we simply have to program one piece of software independent of the provider.

What is the impact?
We need to envisage the situation in a few years, when the amount and complexity of data has been increased in all sciences and there is a greater need to rely on automatic processes, as human intervention means loss of efficiency.
In such scenarios, communities can exploit the wealth of the data domain relying on semantic interoperability between all relevant actors for example for Big Data analytics. The above example is just one small usage scenario that would be enabled if the relevant PID service providers accept the results of the PIT WG and harmonize their approach. Application software writing would be reduced dramatically since only one API would be supported and one module would be sufficient for retrieving the checksum, for example, and checking identity and integrity.

The strengthening of PID information types could also move the existing identifier systems and the overall idea of identification into a more central and fundamental position as suggested by DFT’s core model of a Digital Object, leading to an enormous increase in efficiency when dealing with data.

**When can we use this?**

First groups are building software to implement a first prototype based on the defined PIT API. This first prototype works together with the DTR prototype and both are publicly available, but not designed for production use. We expect another update of the prototypes to become available for download at the end of 2014.

Please check the information on the PIT group’s web-page at [https://www.rd-alliance.org/group/pid-information-types-wg.html](https://www.rd-alliance.org/group/pid-information-types-wg.html)

It is now time to convince the PID service providers to adopt the solution.

Assume that you got a list of PIDs referring to data you want to use in a computation, that these PIDs are being registered at different providers and that you first want to check whether all data objects are still the same. You simply want to provide one module that reads a PID from the list and submits a request to the appropriate resolver to send the checksum. If all actors refer to the same entry in the DTR interoperability is given, i.e. one module would be sufficient to retrieve the checksums independent of the internal terminology used by the various providers.

**What is RDA?**

The Research Data Alliance (RDA) was planned and launched in March 2013 by an international group of collaborating data professionals from research with a vision of researchers and innovators openly sharing data across technologies, disciplines, and countries to address the grand challenges of society. Members of the RDA voluntarily work together, in a self-forming and completely voluntary manner, in self-formed Working groups or exploratory Interest groups. RDA is supported by the European Commission, the United States National Science Foundation and the Australian Government. Information can be found on [www.rd-alliance.org](http://www.rd-alliance.org).

**Produced by RDA Europe**

rda-outputs@europe.rd-alliance.org