PID Policy Directions

Peter

Last week, Karel Luyben presented some slides at a PID meeting in Amsterdam organised by GEDE. His slides, to which several from the FAIR and Architecture groups contributed, include some statements which seem to be broadly agreed. Perhaps they form a good starting point for consensus finding on PID policies. We could compare them with other documents aggregated so far to look for missing statements which should be included.

The slides from Karel can be found here (<https://github.com/GEDE-RDA-Europe/GEDE/tree/master/PID/PID-workshop-Amsterdam>). In case of interest one can also find all other slides presented at the Amsterdam workshop where we had intensive discussions. These resulted in a summary statement called "Key Requirements for a European PID System for Every Researcher" which is now also being circulated in the GEDE group.

I should mention that Karel included the FAIR Digital Objects as a pillar for the EOSC infrastructure building, i.e. as is probably known: PIDs are an essential pillar of the FAIR DOs.

# 1. Karel Luyben Slides

## 1.1 Key Issues (widely agreed)

* EOSC policy needs to be balanced (URIs, Handles, DOIs, etc.) and can't one approach over another.
* We need to address all use cases - data labs and publications
* PIDs need to be globally unique, persistent and resolvable
* Interoperability between PIDs (see below)
* Scalability (trillions of PIDs due to FAIR F1)
* PID system needs to be served by trusted registration authorities and service providers
* Desire for lean, scalable and low-cost PID services
* PID types need to be mature - have stable and trusted schemes with established PID service providers

The last statement was discussed and not understood. I allow me to make a diagram that may help to figure out what is meant. Two paths are normal to come to a PID of a FDO (search or reference found somewhere). They both result in having a (set of) PID. The PID suffix is created according to some scheme by the provider. The PID is then resolved to a structured attribute set the types of which must be defined in a registry to allow machine interpretation. Dependent on the type of FDOs (data, software, assertion, organisation, etc.) and the repository the set of supported attributes (profile) will be different and there must be an API to get the profile being used. The attributes should at least point to the locations where the bit sequences can be found, what the PID of the descriptive/scientific metadata is, what the type is etc. It should be noted that RDA defined the data type registry and defined a core set of PID attributes (which are registered in a registry).

metadata

search

metadata

records

(set of)

PIDs

citation or

reference

structured

attributes

resolved to

include PID

DO's bit

sequence

descriptive

metadata

other

metadata

Type Registry

or Ontology

defined

defined

suffix

scheme

Interoperability between PIDs is a good request, but we need to say what that means. As far as I know there is only one document (ITU Standard X.1255) which describes interoperability in abstract terms. To me "PID interoperability" means

* PIDs to be resolvable to a defined attribute set (profile) of which the attributes are all declared types so that machines can find their way and that we use these types[[1]](#footnote-1)
* following recommendations about attribute profiles for different types of DOs (data, metadata, software, assertions, etc.)

Perhaps, other colleagues have other interoperability aspects in mind. Using the web, we do not need one resolution system, since proxies will b used.

## 1.2 Basic PID Services (debated)

Karel also presented a slide about PID services which was intensively discussed, i.e. different opinions could be observed. While the FDO speaks about a "bit sequence" encoding its content, Karel's slide include the term "resource" which is a simple mapping.

|  |  |  |
| --- | --- | --- |
| Service | Karel's Slide | FAIR DO Position |
| PID creation | creating a PID according to a scheme | creating a PID according to a scheme and a profile with typed attributes (which in general includes for example the cksm |
| PID assignment | assigning a PID scheme to a resource and association of metadata to the resource | assigning a PID to a FDO which is also being described by a metadata FDO |
| PID resolution | direction to the associated resource from a PID under a PID scheme | a PID is resolved to a structured attribute set according to a profile of registered and defined attributes |
| PID description | delivery of the metadata associated with a PID under a PID scheme to allow the user to assess the associated resource | this concept does not exist, since metadata is separated from the PID; its resolution will point to the metadata, but except being machine actionable the FDO concept does not make a statement about metadata - this is left to the knowledgeable communities |
| PID search | PID collection to be searchable via its metadata to discover resources | PID search could be made based on the stored profiles, but does not make much sense; metadata search is separated from the PID |

The difference mainly comes from the fact that PID and metadata are separated. While PIDs can be standardised, metadata descriptions are defined by the scientific communities to foster their scientific work. Generic schemas such as DublinCore are used for generic search, but are useless for scientific work. In fact all scientific communities have developed their metadata schemas which goes so far that they exactly the describe properties of their sensors (instruments), since for certain calculations thy need to know which filter characteristics have been selected.

We see here a difference between the publication use case and the scientific use cases.

## 1.3 Conditions for PID Service Providers (debated)

Karel also presented one slide on conditions for PID Service Providers where only small differences could be identified.

|  |  |
| --- | --- |
| Karel's slide | FDO |
| PSP should have a sustainability plan | same in other words |
| PSP should be governed by a community relevant to EOSC | EU PID system should be governed by the scientific community |
| PSP should make its requirements on repositories clear | not obvious what is meant here; FDO requires uniqueness, profile specification, typed attributes, correct attribute values, etc. |
| PSP should have procedure for change control and versioning | not obvious what is meant here; |
| PSP should have appropriated SLA for reliability | same in other words |
| PSP should have a guarantor for maintaining links if resource disappears or the PSP disappears | standardised tombstone notes are required; if PSP disappears, redundant PSP (mirror) needs to take over |

In general, the requirements are similar, some wording needs to be chosen carefully and some assertions are not clear.

## 1.4 Feedback from Event

Karel also had one slide with feedback from EOSC@RDA Event. Some were discussed in the Amsterdam meeting.

|  |  |
| --- | --- |
| Feedback | FDO |
| machine actionability needed | absolutely - FAIR assumption |
| tombstone when content removed | same - see above |
| granularity & versioning | no general solution, except be as granular as possible; granularity is a choice of the community and/or repository; some repositories use versioning as attributes in the PID record |
| different requirements of PIDs across different stages in the research life cycle | same |
| policy needs to be high level but be accompanied with technical implementation guidance | not addressed but ok |

# 2. Requirements from FDO - Paris/Amsterdam work

Here I list some requirements for a European PID system that come from the FDO domain as being discussed recently and that were not already listed above. These are currently being circulated in GEDE to get more responses.

* be established considering a long-term survival of the referential system (> 100 y)
* support the FAIR DO concept (implicitly included in Luyben's slides)
* offer PID services for all EU researchers in all EU countries (not excluding collaborators from outside of EU etc.)
* focus first on basic services, applying lean organisation principles, all at acceptable costs
* leave the metadata discussions to the domain experts
* offer a high degree of flexibility in some dimensions (branding, suffix schemes, attribute profiles, etc.) without hurting interoperability, i.e. type registration
* control by the EU scientific community (service characteristics, business model) using fair representation mechanisms
* ensure that principles of Open Science are upheld and that monopolisation is prevented
* ensure that crucial registries are independent units under public control
* inclusion of all FDOs from all trustworthy repositories in all kinds of operations
* ensure independence from underlying technology
* built in security from the beginning

# 3. Requirements from other Documents

Here requirements at policy level are being mentioned that have not been addressed by the statements in the previous chapters.

### Requirements from FAIR DO

* Each DO identified by a PID can be accessed or operated on using an interface protocol by specifying the PID of an operation.

### GEDE PID Document

* Address space larger than the maximum number of identifiers required in the foreseeable future.
* A PID needs to be actionable on the web.
* A PID registration and resolution system must ensure 24/7 availability and provide openly documented APIs.

### PLOS Paper

* avoid embedding meaning
* do not reassign or delete identifiers
* document the identifiers used and display responsibility
* PID should ideally be assigned at no costs to individuals depositing data in a repository.

### ODIN Project and Others

* ORCID should be used to identify persons

### FREYA Scoping Notes

* Recommended core PID services ned to be defined and there will also be desirable services for EOSC.
* A discussion of governance of PID services needs to be started considering the global dimension and sustainability.

### RDA Groups

* we should refer to a variety of RDA groups being active.

1. Some argue that we need a type ontology at the end. [↑](#footnote-ref-1)