Response to "Scoping a PID Policy for the EOSC"

Peter, October

Th FREYA project submitted a document to the PID Policy Group in the EOSC FAIR WG. This adds a valuable contribution to the set of documents which we have in the pile of relevant documents uploaded and to be considered by the members of the PID policy sub-group.

In this document I comment on the content of the FREYA contribution.

## General Points

* The points made are very well described and easy to read.
* There are different discussion strands about PIDs at different places and the document indicates to me that we all are not very well integrated and aligned. FREYA is one of them. Other relevant strands are the FAIR-DO initiative, the WEB-experts, the DONA initiative, the data lab experts organised partly in GEDE, RDA discussions, industrial groups, etc.
* This FREYA document motivates to go back to a more abstract level of discussion compared to what has already been done in other strands.

## Statements

1. Statement 1 requires a proper terminology. Indeed a proper terminology is urgently required in the field of research data management. However, just in the area of PIDs much terminology work has been in different discussion strands mentioned above. So we need to build on what has been done over the last 20 years and not start from scratch.

It is requested to associate properties also in relation to digital objects. This is indeed necessary, but there is a broad group in RDA, RDA GEDE and globally that made statements about such properties and that worked out the notion of the FAIR Digital Objects in great detail. This knowledge should be taken up. Again we do not start from scratch here. The GEDE DO includes about 150 experts from many different ESFRIs and also global institutions.

1. Statement 2 requires defining essential classes of DOs to be recommended for EOSC. This is absolutely right. Whole RDA, GEDE and DONA discussions worked on this matter and came to conclusions during the last years. However, these groups speak about DO "types" and they are core for encapsulation and others. FREYA does not make a reference to the RDA DFT, Kernel and DTR groups which discussed this aspect since 2013. The DTR group even suggested and implemented a Data Type Registry which is already being used by various projects globally. DONA worked out the DO Interface Proposal which is fully relying on "types". To me it seems that "classes" and "types" in this context are synonyms.

Another aspect which was agreed in these discussions is that scientific work is dynamic and thus new types will be invented by researchers and need to be accepted, since otherwise we would hamper scientific progress. However, we need to demand that new types are registered in a type registry as defined for example in RDA or in schema registries.

A third aspect was discussed recently in the FAIR DO community. It may be necessary to not only have registries (can be multiple), but build ontologies on top of that to indicate relationships between different types. Some people in GOFAIR are working on such an ontological approach.

1. Statement 3 seems to talk about an issue, which the above mentioned DO community calls "kernel attributes". A first set of such kernel attributes has been defined in RDA and it is agreed that DO providers (repositories, registries) should show their kernel attribute profiles referring to the type registry to allow machine actionability. There are ways how to provide such profiles on request via an API. What does "qualification for EOSC" in the FREYA statement mean? I am not quite sure whether I understand, but in the other discussion strands mentioned it is clear that repositories need some degree of flexibility with respect to the kernel attributes being used to serve their scientific obligations. The principle requirement is that the attributes being used are registered in a type registry to achieve interoperability.
2. Statement 4 asks for "recommended core services" and differentiates "desirable services". This differentiation makes much sense indeed. In the above mentioned discussion strands the first has been settled making a difference between global and local PID service providers. For the Handle System[[1]](#footnote-1), for example, it is a matter of very strict regulations which kinds of services the root nodes participating in the globally distributed root network have to offer, since we are all dependent on a functioning and robust root network. This is being discussed in the DONA Foundation having members from many continents, from ITU and the DOI Foundation.

With respect to the local services there can be a large difference. Currently, there are about 4000 Local Handle Servers in production globally with a large fraction in Europe. In industry the service definitions are different from those in science. For users who want to do a data publication act the services are different from those running labs with for example many sensors producing DOs at a high rate or running a simulation/calculation.

Common to all are the basic services to register PIDs, to resolve PIDs and to manage these PIDs in a secure way. Many added value services on top of PIDs can be thought of such as linking different DOs and DO types, finding duplicates, etc. Key will be to address the scalability issues.

1. Statement 5 speaks about PID Authorities and Service Providers. This statement points indeed to an active discussion which in my view is not settled yet and indeed needs further discussions and practical tests. The big challenge we are faced with is scalability. FAIR Principle F1 requires that data and metadata are assigned a PID. This will lead to trillions of PIDs in the next decade. Some labs already follow this practice and create millions of PIDs (in general Handles) which is possible with the available technology. If now all scientific communities in all countries will "turn FAIR into practice" we need to address the scalability issue. This will require that we are flexible in responding to the way we set up Authorities and SPs.

As already mentioned, in the case of the Global Handle Resolution system the DONA Foundation is taking care that the rules for the root nodes (MPAs) are being followed strictly, that the system evolves smoothly, that the business model is self-sustainable, that no particular company or institution can control the DONA Foundation, etc. This system currently has also some Registration Authorities handing out prefixes which currently are the root nodes, but this model may change over time. The DOI Foundation, being part of the global Handle community with prefix 10 and as such represented in the DONA Board, also has such a model enabling different SPs (such as Crossref and Datacite) to issue Handles based on sub-prefix domains.

Further developments with respect to RAs and SPs need to be observed and measures need to be taken by the relevant initiatives in a coordinated manner. But there should be no one who prevents institutions to setup their own local Handle System, for example, since the requirements are varying and nevertheless integrate it into the global and integrative resolution system.

1. Statement 6 asks for a Governance of PID Systems for EOSC to address sustainability etc. This is indeed a topic that needs to be discussed, but the discussions need to be inclusive, i.e. respect the different usages and respect what has been setup already. Sustainability cannot be solved by asking again and again funds from the EC and other funders. Like DONA and the ePIC service there must be a funding of the basic services by the interested communities. Added value services such as addressed in FREYA and others are important, but normal project funding can be used to set up and test such services, but they need to be turned in self-funded activities over time. Obviously, added-value services will be an area where new teams will work out new services continuously in a competitive domain. Therefore, it is important that we establish an open domain of PIDs to not exclude startups etc.

Certification schemes will be necessary to be able to make statements about the quality of services. DONA, IDF and ePIC, for example, have testing rules in practice, but it would be good if assessments would be done by external institutions based on rule sets eventually being defined by EOSC and to include all local service providers.

## Summary

FREYA addresses a variety of relevant aspects, but their discussions seem to not be linked to the other discussion strands that have an active interchange and that have been driving discussions to solutions for a number of issues raised. I cannot claim that these solutions are all perfect, but I would expect that the FREYA colleagues would know about them, refer to them and eventually criticize them if necessary.

The EOSC FAIR PID Policy Group needs to be integrative and we need to take a "balanced approach", i.e. we cannot ignore, for example, the huge domain of Web people claiming that for example URIs do solve all issues. The whole Linked Open Data community is kind of bound to URIs by the definition from TB Lee. Therefore, in any document from the PID subgroup we need to indicate what the "scope of inclusiveness" is. If we would exclude, for example, URIs, we would have to argue why. The Chinese authorities setting up national PID services for science, industry and public sector, for example, discuss BAR-code solutions which are very common in industry.

Appendix

Scoping a PID Policy for the EOSC

FREYA project

23 Sept 2019

The EOSC FAIR and Architecture WGs are jointly developing recommendations for a PID Policy for the EOSC. The FREYA project is working with the WGs in this task and is developing policy recommendations. As a first step, it is necessary to set the scope of the policy to ensure that the policy covers the appropriate aspects of the use of PIDs within the EOSC, and a discussion can be held on what should be included or excluded from the policy. In this document, we describe those topics that we consider should be in scope of a PID policy; the details of how a policy might be proposed for each topic will then be the subject of a discussion document with detailed recommendations.

We take as a baseline the statements from the Turning FAIR into Reality report that:

* digital objects should be associated with a PID, and
* to support the use of PIDs in the EOSC data infrastructure, PID services should be available.

Thus we need to cover within the PID policy:

1. A definition of terminology for PIDs for use within the EOSC. This would include the properties that we consider as essential to characterise a PID and its association with digital objects within the EOSC.
2. In particular, we would consider which classes of digital objects should be considered as essential for use in a practical working EOSC, and thus should have a PID associated with them, and which should be recommended as suitable for further development to enhance the capability of the EOSC.
3. Together, these would allow us to recommend what properties an issuer of a particular set of PIDs should enforce to qualify as valid PIDs within the EOSC.
4. A definition of recommended core PID services which are essential for the operation of the EOSC, and additionally those PID services that are desirable to enhance the functionality of the EOSC. This should include a consideration of which core PID services should be included within the EOSC Federating Core.
5. Classes of PIDs and associated services are governed by PID Authorities and supplied by PID Service Providers (PSPs). The Policy should include a consideration of what criteria PID Authorities and PSPs should satisfy so that they can be seen as trusted to support suitable PIDs and PIDs services within the needs of the EOSC, and thus qualify as suitable for use within the EOSC.
6. A discussion on the Governance of PID Services within the EOSC structures, especially considering the necessarily global reach of PID infrastructure. This should also consider sustainability actions which the EOSC might undertake to maintain PID infrastructure.

We would expect that the recommendations would be illustrated by typical generic usage scenarios to support the recommended policy choices.

1. In this document I refer to the Handle System, simply because I know it very well. But similar statements could be made with respect to other systems. [↑](#footnote-ref-1)