Group of European Data Experts

Bringing People Together to Advance Data Science!

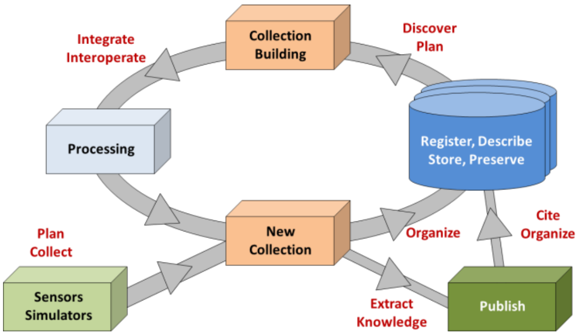
Key Requirements for a European PID System for Every Researcher

- consensus version amongst the workshop participants -

At the Amsterdam Meeting about a European PID System organised by RDA GEDE at 4/5. 11.2019 it was agreed to formulate "key requirements for a European PID system for every researcher" which show rough consensus amongst the workshop participants. In a second step this version will be distributed amongst the GEDE members to get comments and achieve rough consensus within the GEDE collaboration. The participants see the agreed upon requirements as a major contribution towards setting up a balanced PID system within EOSC.

Data Life Cycle

The data life cycle can best be described by a diagram that emerged within the RDA Data Fabric group [1]. It describes the cycle of data creation and processing in the many data labs maintaining some form of repositories and the final act of publishing some results. Here are some key facts that are known about this data life cycle:



- Research organisations maintain many repositories (>1000) of different quality, covering far more than 90% of the scientific data;
- According to Heidorn 80% of the scientific data is "dark data" [2], i.e. nor findable nor accessible[1];
- Some repositories are well-organised and managed, follow certification guidelines and apply PIDs at high granularity to support data science effectively[2];
- Increasingly often, PIDs are used for stable referencing in collections, workflows, etc. to support reproducibility.
- Much of the scientific data is non-FAIR limiting their discovery, reuse thereby restricting their value and the rate of scientific discovery.

[1] Many experts believe that the fraction of "dark data" is much higher.

[2] Mostly Handles are used and at more granular level than DOIs. Most still hesitate to take that step since they lack a suitable service.

Much effort has been put in organising the scientific "publishing domain", driven mainly by the publishers, data librarians and stewards. This domain is focusing on linking different types of digital objects, on calculating metrics and thus creating visibility for researchers and research organisations. Currently, only DOIs are accepted to be part of the mechanisms used in the publishing domain.

An expressed and growing concern is that the ideas of Open Science are undermined by an increasing trend to establish closed domains of "published" results dominated by publisher interests. It is not obvious what exactly the "publication" of scientific data includes or excludes, and why FAIR and persistent references to data in trustworthy repositories would not be acceptable. At the same time, existing publication mechanisms are no longer adequate in the domain of digital data science where all kinds of digital objects at different life cycle states and for different purposes are being referenced and are essential parts of Open Science as planned within EOSC.

It is widely agreed that the current investments in data infrastructure development are so high that its basic design principles and its referential system need to be planned for a long life-time (> 100 years). Only such a long-term perspective can create the trust needed for researchers to participate. PIDs used in scientific research must therefore be able to guarantee persistence over long periods of time as well. This can be best achieved by insuring that the PIDs and their resolution system be implemented in such a manner as to be independent of underlying technologies as much as possible, avoid proprietary technologies, should be free of semantics and free of commercial interests. Given the amount of available and future scientific data and its complexity, data management procedures will need to be automated and foster "machine-actionability" at all levels. Security will have to be built in from the very beginning as they cannot be properly added later.

General Requirements

A European PID System (EUPS) as part of EOSC must:

- Be established considering a long-term survival of the referential system (> 100 years);
- Be machine-actionable as a core principle;
- Address the different use case scenarios that range from modern data intensive science to the more traditional act of data publication;
- Address the (scalability) implication of the FAIR Principle F1, which states that all metadata and data be respectively assigned a globally unique and persistent identifier;
- Support the concept of FAIR Digital Objects (FAIR-DO), which are now accepted as the integrative technology for data infrastructures;
- Offer PID services for all European researchers in all European countries (but also not excluding offering services beyond Europe for those wishing to make use of them);
- Focus first on basic services, applying lean organisation principles and offer them at acceptable costs;
- Leave discussions about specific metadata schemas to the experts in the appropriate communities;
- Offer a high degree of flexibility at some dimensions (branding, suffix schemes, attribute profiles) without hurting interoperability criteria; and,
- Be controlled by the scientific community with respect to service characteristics and business model aspects, whilst respecting the national rules and organisational forms;.

Technical Requirements

A European PID System (EUPS) as part of EOSC must:

- Be based upon the Handle System which includes DOIs;
- Ensure the inclusion of all digital objects from all trustworthy repositories, i.e. it must follow an inclusive strategy;

- Consider that there is much legacy data and that often converting their host repositories may be difficult from a technical or socio-economical perspective;
- Address the scalability challenge of supporting the creation and resolution of trillions of PIDs by defining an appropriate architectural model that i) respects the national and scientific community needs and ii) fosters delegation mechanisms;
- Be ready to integrate all trustworthy actors who share the goals and meet the rule sets;
- Support flexibility with respect to PID record attributes[3] while requesting that those being used be registered according to a defined registration process;
- Ensure that calculated statistics be inclusive;
- Ensure independence from underlying technologies;
- Build in security from the beginning; and,
- Support accredited checking services that make PIDs more valuable (link check, registration of PID record attributes, etc.).

Governance Requirements

A European PID System (EUPS) as part of EOSC must:

- Ensure that the scientific community is guiding EUPS and that core communities are well and equally represented;
- Ensure that the principles of Open Science are upheld and that monopolisation is prevented; and,
- Ensure that crucial registries[4] are independent units under public control, upholding Open Science.

Other Aspects

A European PID System (EUPS) should:

- Foster the development of common tools (linking, collection building, registry support, etc.);
- Help accelerating the process to determine best practices in many still open aspects by making use of Research Data Alliance (RDA) and other similar groups to develop recommendations;
- Foster the building of open source software libraries that facilitate the re-use of code snippets relevant in the realm of PIDs;
- Carry out outreach and engagement; and,
- Give help and support to interested parties.

Participants: Christophe Blanchi, Alex Hardisty, Kostas Koumantaros, Milan Ojstersek, Ulrich Schwardmann, Dieter van Uytvanck, Tobias Weigel, Klaas Wierenga, Peter Wittenburg

[1] https://www.rd-alliance.org/group/data-fabric-ig.html
[2]
https://www.researchgate.net/publication/49175975_Shedding_Light_on_the_Dark_Data_in_the_Long_Tail_of_Science

[3] In the realm of RDA these attributes are called Kernel Information.

[4] Some registries are crucial for the functioning of complex infrastructures such as the PID Prefix Registry, the registry for for PID record attributes (types) and probably others. Registries such as maintained by metadata harvesters are not crucial since the content can be harvested again.