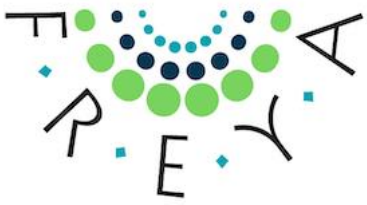


FREYA project workshop for selected EOSC projects

Persistent Identifiers in Research Disciplines

Report on the workshop held on 5 August 2020

Simon Lambert (UKRI-STFC, FREYA project coordinator, simon.lambert@stfc.ac.uk), editor, based on contributions from the individuals named below for each project and summaries of the workshop written by FREYA team members

---

Introduction

FREYA is a project funded by the European Commission's Horizon 2020 programme that is developing the environment for persistent identifiers (PIDs) into a core component of European and global research e-infrastructures. FREYA recognises that PIDs already play an established and vital role in research, but that there are still many opportunities waiting, for example in assigning PIDs to new types of entity, making richer connections between them and enhancing governance for the benefit of stakeholders.

The vision of FREYA is built on three key ideas: the PID Graph, PID Forum and PID Commons. The PID Graph connects and integrates PID systems to create an information map of relationships across PIDs that provides a basis for new services. The PID Forum is a stakeholder community brought together by FREYA, and is active at pidforum.org. The sustainability of the PID infrastructure is the concern of the PID Commons, defining the roles, responsibilities and structures for good self-governance based on consensual decision-making.

FREYA is one of the projects helping to build the European Open Science Cloud (EOSC)—though of course the scope of PIDs is global rather than geographically restricted. FREYA organised a workshop, held on 5 August 2020, whose purpose was to bring together a number of current EOSC-related projects that are devoted to connecting their own research communities to the EOSC, as well as one major project, FAIRsFAIR, which is not discipline-specific but focusses on practical solutions for the use of FAIR data principles. The aim was to take stock of the uses, opportunities and challenges of PIDs across a range of research disciplines and to raise questions such as: what is there in common, where are the differences, what are the opportunities? FREYA itself includes a range of disciplinary partners developing their own applications of the PID Graph, but with the emphasis on the legacy of FREYA in the last part of the project, it was the right time to engage in this concrete way with the current EOSC projects. The outcomes of the workshop will be taken into account in a forthcoming public FREYA deliverable “Integration of the PID Graph with the EOSC”.

Summaries of the project’s presentations and discussion are given below, but it is possible to extract some general points of interest from the PID perspective. These can be divided under three headings.

Diversity in driving forces for PIDs

The need for PIDs may arise from the nature of the discipline itself: for example the timescales over which PIDs are required to persist, with some disciplines needing very long timescales for access to data. The

sheer number of entities requiring PIDs is also very discipline-dependent. Another factor is the practice of research itself in the discipline: for example, the need for reproducibility and provenance (for which PID Graphs can be part of a solution), the enabling of collaborations and workflows, and of course the established behaviours of communities, which might include “home grown” PIDs, not used elsewhere, that satisfy the community's needs in a particular area.

Consequent requirements on PIDs

In addition to their basic function of identifying entities (and more specifically resolving to them), PIDs have a role in interoperability and unification of resources (an example being the identifiers.org resolution service). Scalability is of crucial importance in some domains, which implies a need for automated processing (possibly with Artificial Intelligence techniques). The persistence is also an issue, including clarity about the distinction between the persistence of the identification and of the digital object itself, which are not the same thing.

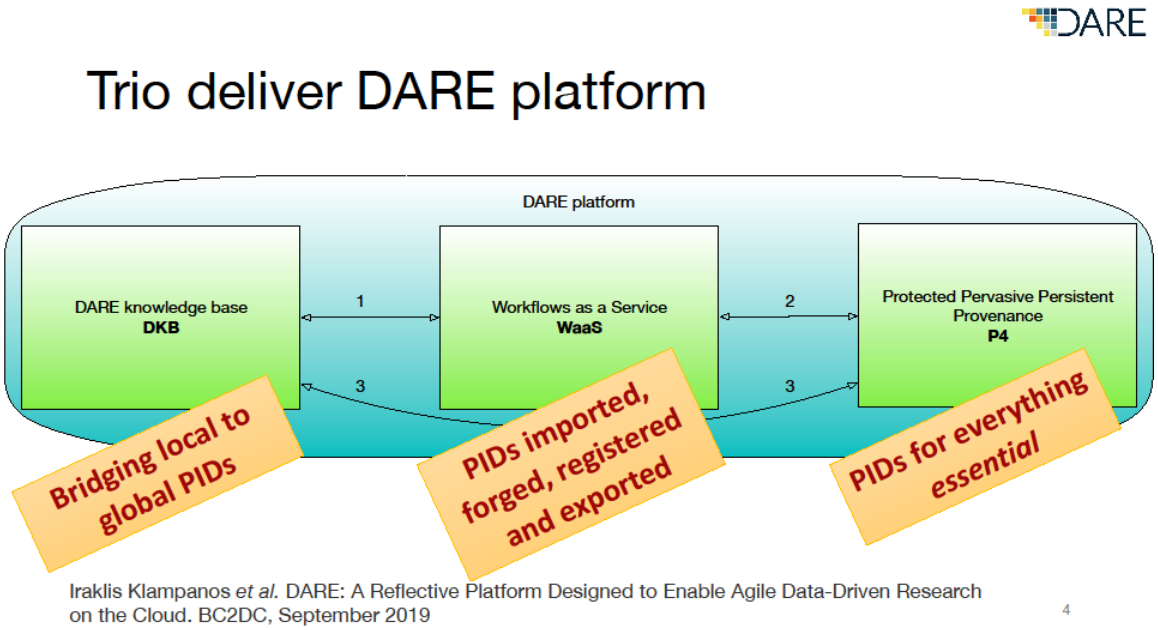
PIDs as part of an infrastructure

Persistence is not an attribute that is acquired without effort: the PID infrastructure must give some assurance through its own sustainability. For PIDs to be effective and attractive in disciplinary use, there needs to be agreement, understanding and trust in what is being provided. Roles in the PID infrastructure are numerous (PID authority, PID service provider, PID manager, PID owner, PID end user). Although the landscape of PID provision seems well established now, there might be disruptions in future, perhaps through commercial competition.

DARE (Malcolm Atkinson, University of Edinburgh)

project-dare.eu

The DARE project responds to the growing needs of research that is both data-intensive and computation-intensive. Large overlapping collaborations are typical of such research, and DARE works with two demanding application domains, seismology and climate impact modelling. Because of the complexity of the research procedures, organised and reproducible workflows are of great importance, and PIDs should be an integral part of these. Related to this, provenance is also important, and a number of requirements for PIDs follow, including verification of research procedures for reproducibility.



The DARE platform provides tools to develop and execute experiments in terms of workflows, and comprises three components: DARE Knowledge Base, Workflows as a Service, Protected Pervasive Persistent Provenance. There are correspondingly different needs for PIDs but they are everywhere pervasive.

The timescales of the research might be long, potentially many decades, and the implementation requires sustainable PIDs for all the entities relevant in establishing provenance over such timescales. However, there is also a need for short-lived identifiers at certain stages, which could be promoted to long-lived forms if necessary.

Since the aim is to support multiple research communities, the practices and expectations of individual communities must be met, yet there should be some uniformity of policies and PID descriptions. At the highest level, PIDs are divided into two groups: system PIDs, relating to the mapping of the research and production work; and application PIDs, relating to the entities of interest to the application discipline.

ELIXIR (Sirarat Sarntivijai, ELIXIR consortium)

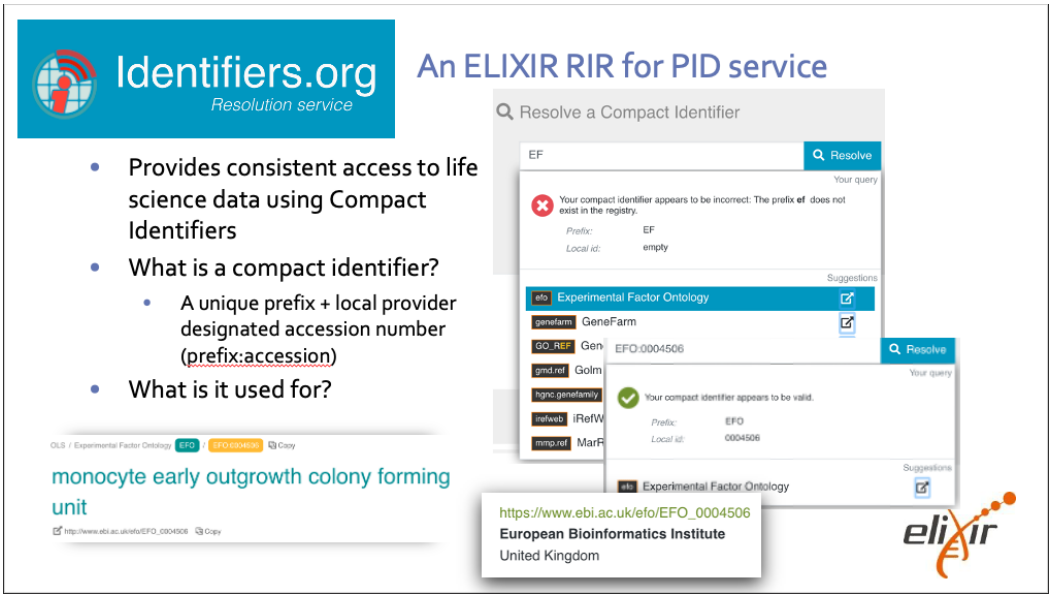
www.elixir-europe.eu

ELIXIR focuses mainly on Life Sciences, connecting computational infrastructures and data between 23 member countries. With a portfolio of five different platforms (Compute, Data, Tools, Training and Interoperability), the workshop presentation focussed on ELIXIR's Interoperability platform (EIP). As for FREYA activities, this is use-case driven and PIDs play a key role.

The EIP Service framework relies on three pillars for linking data and functionality, in a FAIR context:

- Metadata annotations, driven by Bioschemas
- Metadata Services and Standards Registry
- Identifiers and Mapping Services

The contribution to the COVID-19 data portal earlier this year was a good test of how interoperable the platform is, in aggregating data from different sources. Two specific components of the EIP that are especially noteworthy for encouraging participation by different resources and for capacity building are Bioschemas.org (to ensure that a common language of metadata is used for PIDs) and ELIXIR Recommended Interoperability Resources (RIRs). An example of one such RIR is Identifiers.org to ensure consistent access to life science data using Compact Identifiers, a unique "prefix + local accession" combination.



ELIXIR promotes the adoption of PIDs by using them and demonstrating that they underlie FAIR implementation - the EIP Service Reference Framework clearly includes a series of PID centric activities e.g. identifier minting by PID providers, citation (and promoting PIDs to data consumers and producers for this purpose), identifier mapping, and identifier resolution. A core task of the EIP is to deliver cloud-enabled RIRs and services to support EOSC projects.

Training needs are recognized as key to capacity building for interoperability: The platform offers workshops on EIP services and practices. More broadly, training needs are addressed via ELIXIR's training platform. TeSS is the training registry for the ELIXIR platform. Both sites include searchable content and webinars on identifiers.

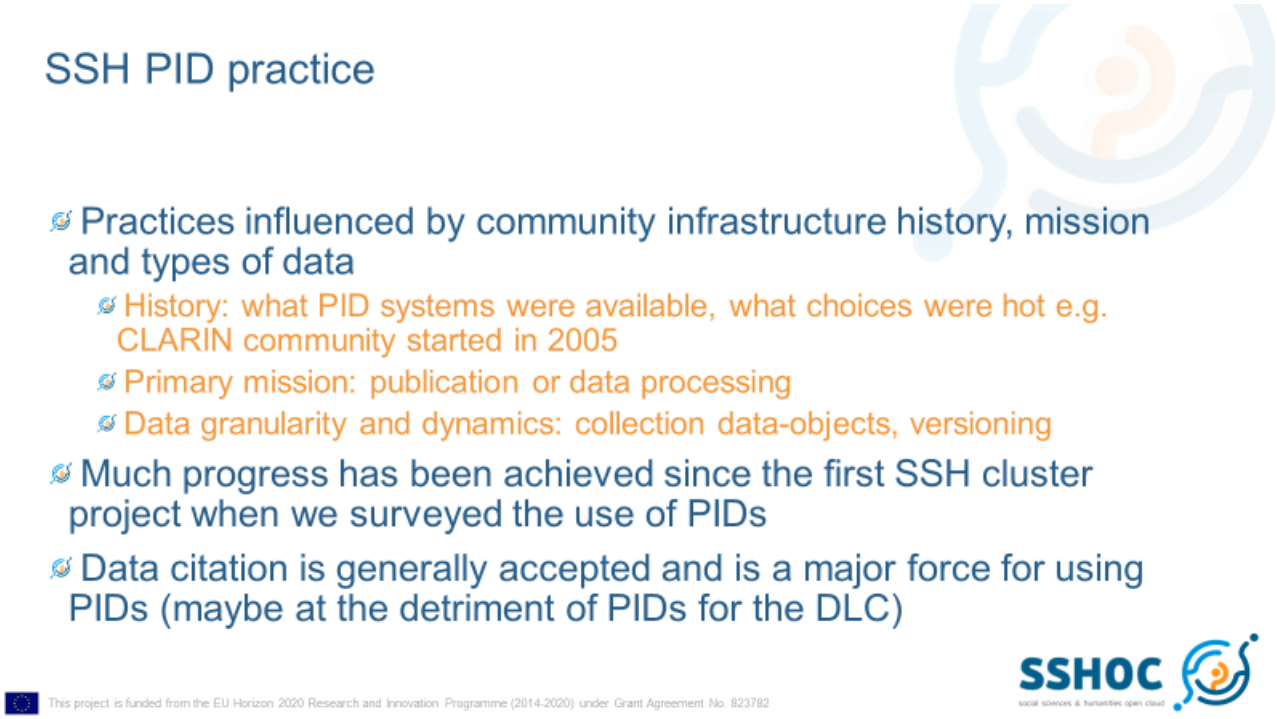
In summary: the thinking on PIDs within this community is mature, albeit focussed on biomedical data as a resource. There would be a good community to test the appetite for extending FREYA's PID graph building initiatives.

SSHOC (Daan Broeder, KNAW/HuC, CLARIN ERIC, and Nicolas Larrousse, Huma-Num – CNRS)

www.sshopencloud.eu

The SSHOC (Social Sciences and Humanities Open Cloud) project is the EOSC cluster project focused on the Social Sciences and Humanities (SSH). The project follows up on earlier initiatives to bring together the Social Sciences and Humanities such as DASISH, PARTHENOS and SERISS. The SSHOC project includes major SSHOC stakeholders including the SSH ERICs CESSDA, CLARIN, DARIAH, ESS and SHARE.

Practices around PIDs in SSH are often influenced by the history of a given field and the available infrastructure. In general, however, the use of PIDs has progressed since the earlier cluster projects. The ERICs generally have or are in the process of finalizing PID policies. There is awareness and involvement of RDA initiatives and recommendations related to PIDs and SSH partners are involved with DataCite.



The use of PIDs in particular for data citation is accepted in the SSH community. However, there is still a lack of awareness of the function and working of PIDs. For instance, the presence of a PID may be seen as a measure of data quality by researchers. The prestige of a service such as DataCite's DOI is assumed to ensure a high data quality indicating a misunderstanding of the relationship between the PID providers and the data providers. Similarly, there may also be confusion between persistence of data and persistence of the PID itself.

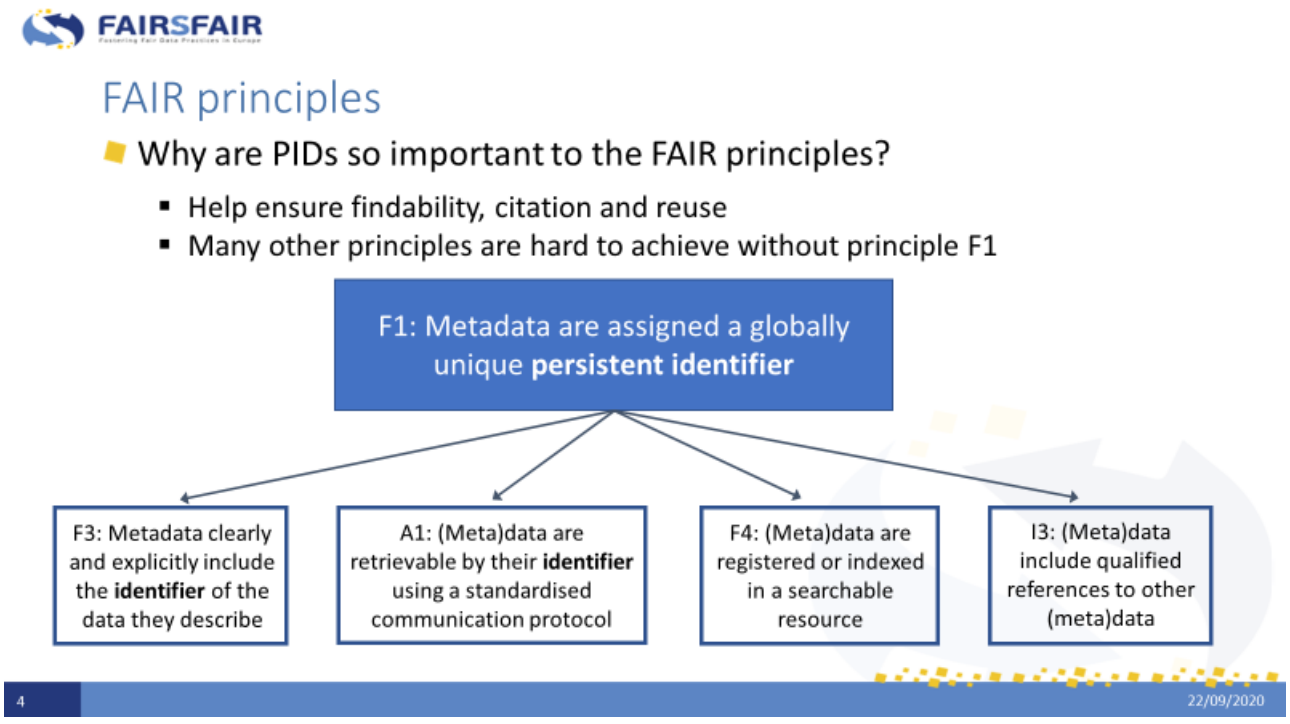
In the SSHOC project itself, the work around PIDs is focused on PIDs in the context of data citations. The project is working on a demonstrator for a FAIR citation repository federation dedicated to the SSH community. Through the PIDs of an object, aggregated information from different sources can be gathered and information can then be enriched through (semantic) annotations. The main goal is to make these standardized and enriched citations machine actionable through a common API in order to foster visibility of SSH resources.

Some of the challenges put forward from the perspective of the SSHOC project include the sustainability of developed infrastructure as well as the connection with other initiatives and developments, thereby mirroring challenges identified in FREYA.

FAIRsFAIR (Jessica Parland von Essen, CSC)

www.fairsfair.eu

FAIRsFAIR (Fostering FAIR Data Practices In Europe) is a large EU Horizon 2020 project with 22 partners from eight EU member states. The project aims to provide practical solutions for using the FAIR data principles throughout the research data life cycle. The project is involved in development of standards, policies and practices as well as community-building and uptake to provide a platform for using and implementing the FAIR principles in the day-to-day work of European researchers, data providers and repositories.



As PIDs play a crucial role in various aspects of the FAIR principles, there is a close link between the work FREYA is doing and the work of FAIRsFAIR. FAIRsFAIR is concerned with the bigger picture where PIDs play a part in improving access to research information as well as aiding research data management and reproducibility.

Of particular relevance for FREYA is the work of FAIRsFAIR work package 2 (WP2) on FAIR Practices: Semantics, Interoperability, and Services which looks at technologies supporting semantic interoperability and aims to develop practices that support FAIRness.

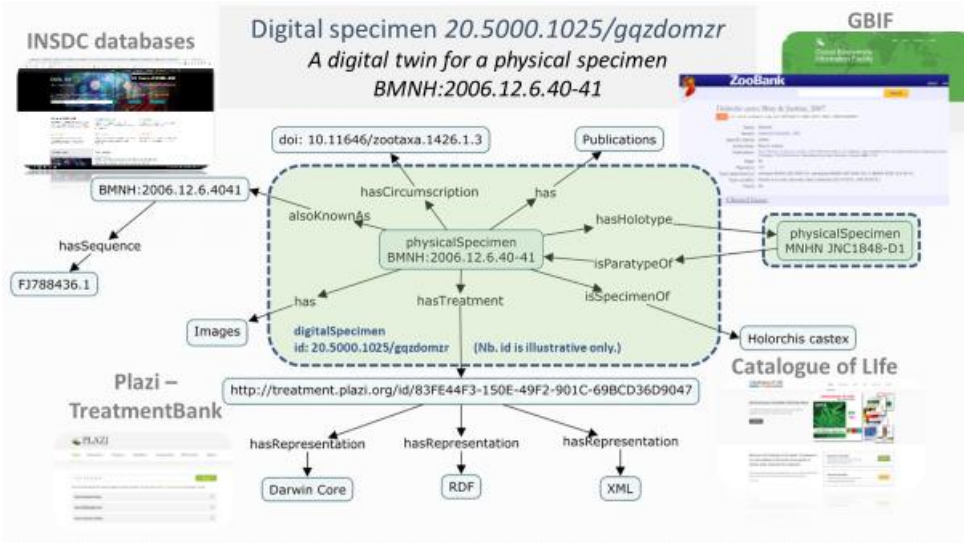
A functioning PIDs infrastructure is necessary to increase the FAIRness of data and the relationship between PIDs and semantic interoperability is very close. There are different roles involved in the PID infrastructure, from PID authorities to the PID service providers, as well as PID manager and owners all the way to the PID user, who all play a part in optimizing the use of PIDs. The interoperability work of FAIRsFAIR targets researchers and service providers in particular. FAIRsFAIR WP2 has produced a second report on persistence and interoperability (Deliverable 2.4) available at https://10.5281/zenodo.4001631. The report is meant to provide examples and guide researchers, data stewards as well as service providers on the use of PIDs, metadata and semantic interoperability. The report puts forward four conclusions:

- A generic solution to achieving FAIRness does not exist but rather decisions need to be tailored, starting from the user's needs.
- Efforts to implement FAIR need to balance the investment and the expected benefits to the scientific community.
- Researchers, data stewards and service providers need to work together to achieve a FAIR data ecosystem.
- Achieving interoperability for both humans and machines requires not only technical solutions, but also human efforts to mitigate misunderstandings and create common understanding of concepts, agreeing on terms and vocabularies and building cohesion based on existing frameworks.

DiSSCo (Alex Hardisty, Cardiff University)

www.dissco.eu

DiSSCo (Distributed System of Scientific Collections) is an initiative of more than 120 national facilities across 21 countries; the largest ever collaboration in the natural sciences collections and biodiversity community. The DiSSCo programme aims to create one European collection of scientific assets, to unify policies, practices and processes, and to provide a mechanism for identifying Digital Specimens on the Internet. Digital Specimens are curated and authoritative packages of links to data assets associated with and/or derived from the physical specimens, such as literature, genomic data, biochemical data, ecological data and taxonomic information. At present it is hugely challenging to make connections across the different data sources which hold the different types of data. This community also operates at large scale and needs very large numbers of identifiers for its collections' objects and associated digital transactions (digital curation, mining, processing, analysis, etc).



DiSSCo has identified the following requirements for a PID scheme:

- Scalability in respect of the specimens, for machines to use them globally. It is estimated that there are 1.5 billion specimens in European public institutions alone. Roughly 2–3 million PIDs per day would be created while DiSSCo is fully operational.
- Appropriate to the situation: a range of identifier types is needed
- Trust and user confidence
- Persistence, potentially over 100 years
- Governance by stakeholders themselves
- Potential for global adoption/expansion

DiSSCo has adopted the approach of Digital Object Architecture and FAIR Digital Objects and that leads to the use of Handles. They reviewed existing Handle PID schemes and decided it might be necessary to create a new identifier brand to establish confidence that the identified Digital Specimen "twin" and its package of links represents a permanent link to the physical specimen from which the data derives. They have proposed a new "Natural Science Identifier" (NSId) and are working on a business case analysis to assess the best way to proceed with that. They are evaluating different routes such as becoming a registration agency or allying with an existing registration agency. DiSSCo will also integrate with other types of identifier, and have noted several emerging PID types including ROR as necessary to their implementation.

It is clear that the thinking on PIDs within this community is relatively mature but that scalability remains a major issue due to the volume of specimens requiring identifiers. As a community they wish to utilise other appropriate identifiers and connect with them, which could be facilitated via the PID Graph. As the NSId specification develops it is important for there to be adequate support for all parties to have the necessary support to develop the identifier and integrate with existing PID infrastructure.

The timescale over which DiSSCo wishes to guarantee persistence is much longer than for other areas, reflecting the origin of the DiSSCo research infrastructure in heritage organisations i.e., natural history museums. Therefore, there are long-term sustainability considerations.

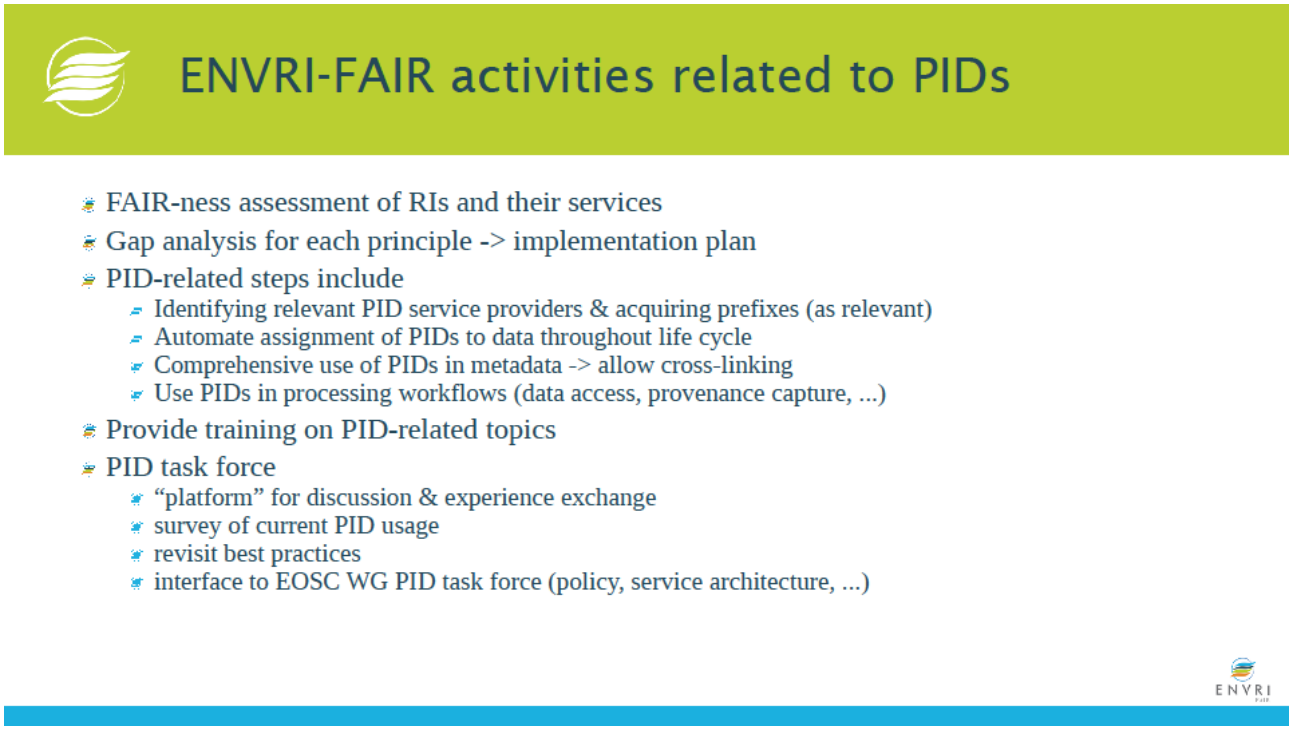
ENVRI-FAIR (Margareta Hellström, ICOS Carbon Portal & Lund University)

https://envri.eu/home-envri-fair/

The ENVRI community consists of European research infrastructures and similar organisations from the environmental and earth science disciplines. It was first conceived around ten years ago, and has steadily grown since then. A series of EC-funded projects have brought together different constellations of ENVRIs working to identify common challenges and solutions, and ENVRI-FAIR is the most recent of these (2019-2022). The participating Research Infrastructures will build a set of FAIR data services, and the cluster will be connected to EOSC.

In the preceding project ENVRIplus, relevant "best practices" for identification and citation of data, as well as other research resources were identified. There was also a study of use cases on e.g. identifiers for data publication, collection of data usage and impact statistics, and strategies for ensuring proper acknowledgement of individual items in data collections.

ENVRI-FAIR partner research infrastructures are very diverse in terms of application domains, maturity level and existing practices and services. ENVRI-FAIR aims to find and apply technological solutions towards making data and services as FAIR as possible. Here the focus is on improving practices in the ENVRI sub-domains (atmosphere, ecosystem and biodiversity, marine and solid earth). Self-assessments of FAIRness have identified a number of gaps, also with respect to PID usage, which are now being addressed. A domain cross-cutting Task Force on PID issues has been set up.



PID-related steps include:

- Identifying relevant PID service providers and acquiring prefixes (as relevant)
- Automating assignment of PIDs to data throughout life cycle
- Comprehensive use of PIDs in metadata to allow cross-linking
- Using PIDs in processing workflows (data access, provenance capture, ...)

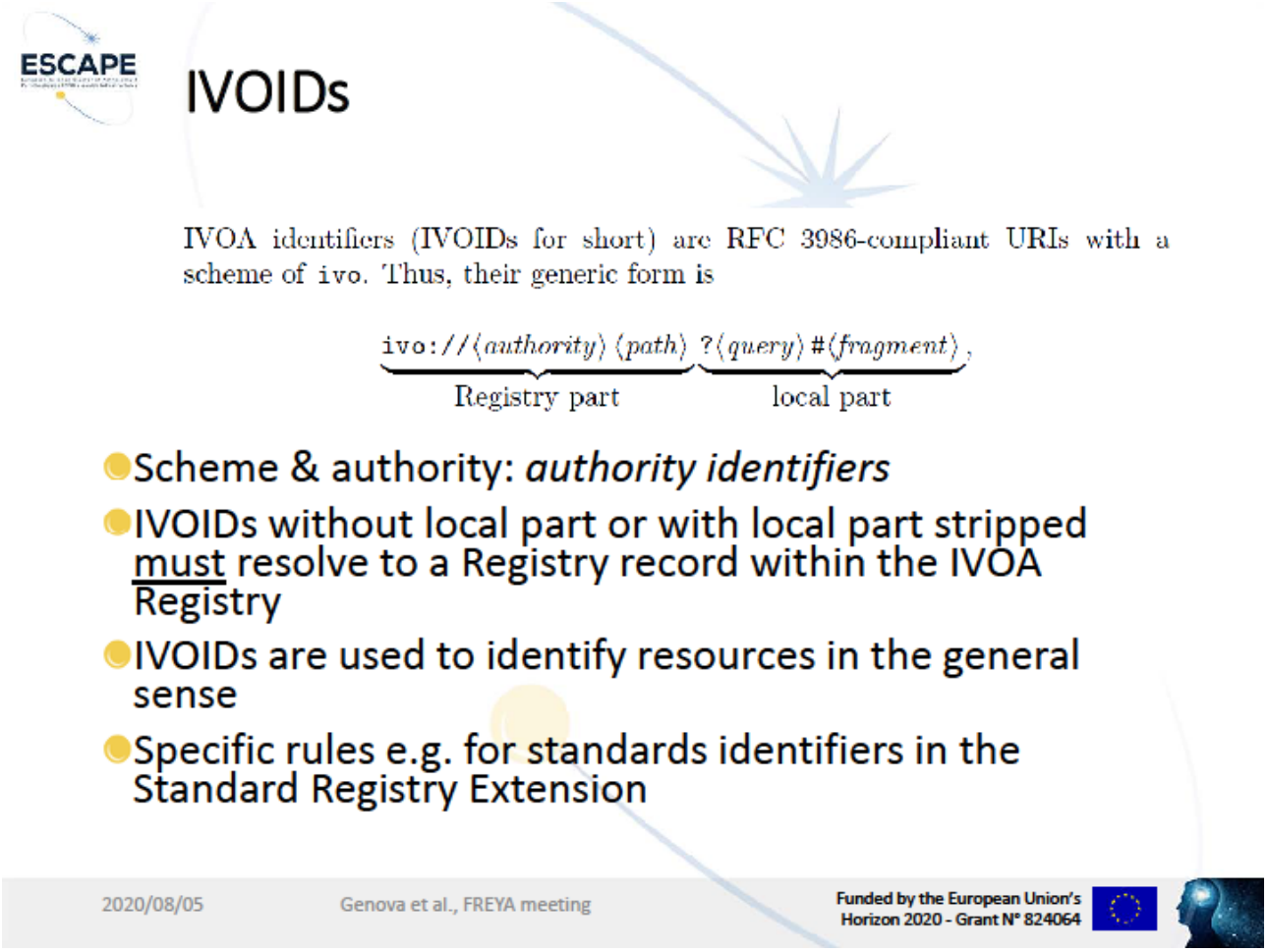
Ultimately, ENVRI-FAIR aims to develop an "ENVRI data hub" for EOSC. Ensuring FAIRness will require easy access to PID services, which could be provided by the EPSC infrastructure, or external providers, raising important questions about the relation of EOSC to the wider world.

ESCAPE (Françoise Genova, Observatoire astronomique de Strasbourg – CNRS)

projectescape.eu

ESCAPE brings together the astronomy and particle physics communities. These communities have some common elements when it comes to approaches for data discovery and PID use. Both astronomy and particle physics are data-intensive fields with well-established Open Science and FAIR practices. The IVOA Data Curation and Preservation (DCP) Interest Group addresses similar topics as the Data Preservation in High Energy Physics (DPHEP) Collaboration.

Various PIDs have been used for years now in both disciplines—ORCID iDs and DOIs are in a mature stage at CERN. PID practice in both cases seems to follow a bottom-up approach in that the first priority is to address the needs of the individual communities. In the same way that IVOIDs fulfil the needs of the community in astronomy and are used in conjunction with other more "globally recognised" and persistent PIDs (i.e. DOIs), similarly at CERN certain information services also make use of non-permanent or internal identifiers that have been developed for community use, while in parallel taking advantage of the benefits of DOIs (e.g. the analysis PID in CERN Analysis Preservation—see FREYA's report D4.1 for more details—or other collaboration database identifiers).



In both cases, PIDs are not necessarily the most common way for researchers to discover data. The astronomical Virtual Observatory is used by the community in its daily research work to access the distributed data resources of the global astronomical data infrastructure. Data shared in the Virtual Observatory can be discovered using many different parameters. In particle physics internal databases and services like INSPIRE or the CERN Open Data portal are used daily by physicists for their work and PIDs are just one component of those systems, which is part of the reason why PID adoption has been challenging in the past in some cases.