Project Name	**FREYA**
Project Title	**Connected Open Identifiers for Discovery, Access and Use of Research Resources**
EC Grant Agreement No	**777523**

D4.5 Integration of the PID Graph with the EOSC

Deliverable type	Report
Dissemination level	Public
Due date	31 May 2020
Authors	Simon Lambert (STFC) Contributions from representatives of EOSC cluster projects and summaries of the workshop for cluster projects written by FREYA team members
Abstract	Review of the state of EOSC and implications for integration with the PID Graph; relationship of key FREYA outputs to EOSC; engagement with EOSC cluster projects.
Status	Submitted to EC 20 December 2020

The FREYA project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 777523.

FREYA project summary

The FREYA project iteratively extends a robust environment for Persistent Identifiers (PIDs) into a core component of European and global research e-infrastructures. The resulting FREYA services will cover a wide range of resources in the research and innovation landscape and enhance the links between them so that they can be exploited in many disciplines and research processes. This will provide an essential building block of the European Open Science Cloud (EOSC). Moreover, the FREYA project will establish an open, sustainable, and trusted framework for collaborative self-governance of PIDs and services built on them.

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, whose members collectively oversee the development and deployment of new PID types; it will be strongly linked to the Research Data Alliance (RDA). The sustainability of the PID infrastructure resulting from FREYA beyond the lifetime of the project itself is the concern of the PID Commons, defining the roles, responsibilities and structures for good self-governance based on consensual decision-making.

The FREYA project builds on the success of the preceding THOR project and involves twelve partner organisations from across the globe, representing PID infrastructure providers and developers, users of PIDs in a wide range of research fields, and publishers.

For more information, visit www.project-freya.eu or email info@project-freya.eu.

Disclaimer

Copyright Notice

Executive summary

The FREYA project has a specific task concerned with integrating the PID Graph with EOSC. EOSC has been in a state of constant evolution and achieving integration is not a simple matter. By thinking of integration as “embedding” it is possible to distinguish different levels and to examine how the results of FREYA are present in EOSC and what still remains to be done.

This deliverable begins by considering what it means to integrate the PID Graph with the EOSC, and how that can be achieved. It goes on to select and examine a number of recent key EOSC documents reflecting the current position from high-level concept down to technical architecture, and to understand the implications for integration with the PID Graph, which might be explicitly referenced or only implicit. Then the key FREYA outputs of relevance to EOSC are identified and their state of integration is examined, using the four levels of interoperability of the European Interoperability Framework as a reference.

A specific activity in this task was to engage with the current EOSC cluster projects from different disciplines, and the deliverable reports on their very valuable perspectives on PIDs in the context of their established infrastructures. Finally some recommendations are given for continuing the process of integration into the future of EOSC.

Contents

1 Introduction and context

Of the three pillars of FREYA—the PID Graph, PID Forum and PID Commons—it is the PID Graph that is the most multifarious. It built upon the developments of the predecessor projects ODIN and THOR, and the existing PID infrastructure; it was inspired by a vision of what could be done with a world of connected persistent identifiers; it supported, and its value was demonstrated through the project's own pilot applications in a wide range of areas of research and with very different priorities. Unlike the PID Forum and the PID Commons, which are described in dedicated deliverables summarising the achievements at the end of the project[1], for the PID Graph a number of deliverables have been produced t the end of the project presenting and positioning the Graph in different ways with different emphases and analyses. This reflects the complexity of the context in which FREYA has been operating for the last three years. This cluster of deliverables all take the PID Graph-related outcomes as they are, rather than produce additional results, though this does not imply that no original work was entailed in their production.

These deliverables produced at the end of FREYA—of which the present report is one—are differentiated by what they place in the foreground.

- **D4.5, "Integration of the PID Graph with the EOSC"** foregrounds those outputs of FREYA that contribute in some way to the European Open Science Cloud.
- **D4.7, "Using the PID Graph: Community Workflows and Discoverability Services"** foregrounds the pilot applications, summarising their final state and the lessons learnt from their development.
- **D5.10, "Dissemination and Exploitation Plan"** foregrounds those outputs of FREYA that are to be exploited by the project partners beyond the end of the project.

It is crucial to position FREYA's outputs with respect to EOSC. FREYA has always existed in the context of EOSC—unlike its predecessor projects ODIN and THOR—and has had envisaged a role in helping to build the EOSC. The very first sentence of the project's objectives declares that "The overarching goal of the FREYA consortium is to iteratively extend a robust environment for a wide range of Persistent Identifiers (PIDs), as an essential component of the European Open Science Cloud (EOSC)." The Description of Work acknowledges that "PIDs, and FREYA, play a key supporting role in enabling the EOSC mission." A number of concrete contributions to EOSC were envisaged at that stage, including integration of services into the EOSC services catalogue; development of PIDs for resources that EOSC will rely on; and pilot applications to show the value of PID infrastructure through the vision of the PID Graph. Most specifically, there is a particular task, of which this deliverable is the outcome, to "Integrate EOSC with the PID Graph".

We understand EOSC to mean not only the general-purpose infrastructure being developed, and the governance and policy arrangements underpinning it, but also the use of that infrastructure to connect established communities to the EOSC. This includes the various community-led projects that have existed and continue to exist, funded by the EOSC-related work programmes of Horizon 2020, initially labelled as "science demonstrators" and currently as "cluster projects"[2]. The FREYA Description of Work refers to the integration activity as concerned with what is needed to support these projects in their adoption of the PID Graph vision and technologies, and that certainly continues to be a valid and worthwhile aim. However, this deliverable is an opportunity to go further than that, and to analyse and position all the PID Graph-related results of FREYA in the context of EOSC. The engagement with cluster projects is part of that: they provide a disciplinary, user-motivated context that is complementary to the high-level EOSC generalities.

In consequence, what this deliverable sets out to do is to explain the position that FREYA has reached, analysing its outputs in terms of how they can be adopted by EOSC stakeholders, and what FREYA contributes to EOSC infrastructure that will be usable in future. In this respect it is in fact closely allied to

[1] PID Forum:D5.7 "Third Report on the PID Forum"; PID Commons: D6.5 "Final Report on PID Commons and Sustainability"

[2] https://www.eosc-portal.eu/news/five-new-esfri-cluster-projects-eosc-panorama introduces the five cluster projects.

sustainability; the work reported here present an angle on the PID Commons through the lens of EOSC specifically. The project outcomes considered include all relevant work done in FREYA including the project's own pilot applications, which could lead to significant impact and attract related communities and stakeholders' interest, as well as the general PID Graph and results of work with EOSC stakeholders on policies and other issues.

This deliverable was planned to appear six months before the end of the project, but in fact has been produced, along with the other "summing up" deliverables listed above, as one of the final outputs of the project, i.e. at Month 36. The reason for the delay was partly this precise fact: that the deliverable is one of several with the same basic function in the project, concluding the work and therefore appropriately produced at the end; but in addition there was the need to take account of the constantly evolving EOSC environment, rather than arbitrarily cutting off six months before it was necessary to do so.

The deliverable is structure in the following way:

- Section 2 considers what it means to integrate the PID Graph with the EOSC, and how that can be achieved.
- Section 3 selects a number of recent key EOSC documents to capture the current thinking on the "whole EOSC" and draw out its implications for integration with the PID Graph.
- Section 4 identifies key FREYA outputs and considers their state of integration with EOSC.
- Section 5 describes FREYA's engagement with the EOSC cluster projects and their particular practices and perspectives on PIDs.
- Section 6 looks to the future with next steps to ensure the continued integration of the PID Graph and other FREYA outcomes with EOSC.

2 Integration with EOSC: what does it mean and how to achieve it?

The task from which this deliverable originates is entitled "Integrate EOSC with the PID Graph", yet the deliverable itself is called "Integration of the PID Graph with the EOSC". The inverted ordering is not carelessness but is in fact significant, reflecting the two possible directions of interaction between FREYA and EOSC. The integration, in whatever form it may take, can have two purposes, depending on whether EOSC or FREYA is taken as the primary focus:

- enhancing uptake, capabilities or support for EOSC through the PID Graph;
- enhancing uptake, capabilities or support for the PID Graph through EOSC.

Given the ever-present emphasis of FREYA on contributing to building the EOSC, the first of these is not surprising; but the second should not be neglected. The PID Graph is a motivating vision in its own right, summarised in the Description of Work thus: "The PID Graph connects and integrates PID systems, representing a map of the relationships across a network of PIDs and serving as a basis for new services." Anything that can advance and support these connections will help to make the vision a reality, and that includes the unifying vision and implementation of EOSC.

The PID Graph has been conceived at three levels:

- a **vision** of linking together a wide range of entities with PIDs to allow complex queries;
- **enabling technologies** to make the vision a reality;
- **applications** showing the vision in reality in diverse domains.

Each of these levels is capable of integration in different ways with the EOSC—which itself is of course a multi-level system.

It is important to understand the meaning of the word "integration" in this context. It is not restricted to technical integration of components of an information system, making them operate together and pass data between themselves as a part of a whole. In fact, the intended meaning could be better expressed as "embedding" at different levels: embedding the FREYA outputs (and particularly those of the PID Graph) into EOSC and vice versa. Embedding means reaching a state of mutual understanding, acceptance and readiness for the concepts, approaches and technologies of the other party, making them available, and extending to technical integration where possible and appropriate.

With this interpretation of integration, examples include making PID Graph related services available in the EOSC Marketplace, and giving recommendations for EOSC actors engaged in PID activities, based on the experiences of FREYA. These are the "knowledge-based resources" referred to in the description of the task. This deliverable aims to show how the embedding of FREYA outcomes has been achieved. In order to do that, the first step is to review the state and recent developments of EOSC, including its plans for the future, and understand their implications for persistent identifiers and their infrastructure.

3 The EOSC landscape and PIDs

EOSC is in a state of continuous evolution. The first wave of EOSC-building projects (including FREYA) is coming to an end, and new ones are starting. The "second phase" of EOSC is approaching in 2021, and a major development has been the establishment of the EOSC Association on 29 July 2020 as a not-for-profit international association based in Belgium (AISBL)[3]. This body will coordinate EOSC-related activities and have a role in definition of policies and adoption of standards. A stream of reports is being issued and events are taking place: the web page for the EOSC Governance Symposium in October 2020 listed nine recent reports as "essential reading" as well as a pointer to the latest version of the Strategic Research and Innovation Agenda (SRIA) for EOSC.

This poses a problem for any attempt to integrate with EOSC—even if the present deliverable is up to date at the time of its finalisation, it will not stay so as EOSC continue to develop. This is recognised as an issue for EOSC in the "Report on the First Synchronisation Force Workshop" of the FAIRsFAIR project (examined below), which records a comment that "What we do in certification and PID policies has to align with rules of participation"—with an implication that such alignment is not yet guaranteed.

In these circumstances, what can be done, starting with the general understanding of EOSC built up over the three years of FREYA from many reports, discussions and events, is to select some important recent contributions to or reflections of the EOSC landscape at different levels, from conceptualisation down to architecture and the EOSC marketplace, pay attention to what is fundamental and the implications for PIDs and their infrastructure, and relate this to the PID Graph outputs of FREYA. In some cases this will reveal that FREYA's outputs have a clear role to play and can be said to be already integrated (in the sense of "embedded" as explained above). In other cases there is still uncertainty and evolution, and all that can be done is to give some general indications for what form eventual integration might take. And it is important not to forget the two-way sense of integration already introduced: support for EOSC through the PID Graph, and support for the PID Graph through EOSC.

The following sections make just such a selection. The criteria for inclusion were completeness of coverage of the EOSC, from high-level definition to architecture, and recent publication, reflecting the position at the end of the FREYA project so as to be as up to date as possible. There are many other documents that might have been included, but a law of diminishing returns would have applied, and it seemed that the documents selected below capture what EOSC is all about at the end of 2020.

3.1 Shared definition of EOSC

Base document

EOSCsecretariat project: "Towards a Shared EOSC Definition" (May 2020)[4]

Context

The complexities and evolutionary path of EOSC have meant that there is little common understanding or definition of what EOSC really is. The EOSCsecretariat project examined a number of authoritative EOSC-related documents from 2015 onwards to try to find characteristic properties of the EOSC.

[3] https://www.eosc-portal.eu/news/eosc-association-registered

[4] https://www.eoscsecretariat.eu/eosc-glossary/post/towards-shared-eosc-definition with additional glossary of terms (latest version at https://www.eoscsecretariat.eu/eosc-glossary/post/second-intermediate-version-eosc-glossary-released)

Summary

The study is interesting for being descriptive rather than normative: it examined actual statements and usages rather then inventing a new definition of EOSC. It built a comprehensive glossary of EOSC-related terms that is "is intended as a basis for the terminological standardisation process within the context of the European Open Science Cloud (EOSC) initiative, with the objective of facilitating and improving the actual communication between the different communities involved."

Overall, the analysis showed that EOSC can be characterised a *system* comprising actors, services, data, policies, and infrastructures, implemented in a federated way and providing seamless access to data and interoperable services supporting the whole research data cycle.

The glossary is a large and very miscellaneous collection of terms, some of which are EOSC-specific (for example "EOSC-Exchange", EOSC-Partnership") but most are not (ranging from "actor" and "administrative metadata" to "trusted digital repository" and "user").

Implications for integration of the PID Graph

Curiously, the term "persistent identifier" (or any of its cognates) does not appear in the current version of the glossary—though PIDs are recognised as "first class citizens" by many EOSC actors. The only mention is under the entry for "digital object" which is defined as "any set of data and an associated unique identifier". Certainly, the glossary is a working draft, but this omission must surely be rectified in future versions. It would also be good to see the term "PID Graph" itself appearing—as noted below under SRIA, it is recognised and used in EOSC discussions.

More generally, considering the characterisation of the EOSC as a system, the attributes of being "trusted" and "providing seamless access" depend on a PID infrastructure and, for their full realisation, a PID Graph that "can also provide unambiguous linking to other resources, both other publications, datasets or software, but also actors such as researchers, research organizations, or funders." The integration here is an underpinning of the some of the desired attributes of EOSC through a vision and enabling technologies of the PID Graph. The Strategic Research and Innovation Agenda of EOSC should take this further.

3.2 Creation and future of EOSC

Base document (1)

Report on the workshop "Co-creating the EOSC: Needs and requirements for future research environments" (March 2020)[5]

Context

A workshop was held in Austria in March 2020 bringing together top-level researchers from different disciplines with a view to develop a vision of how research will be conducted in future, and to contribute directly to the work of the EOSC through its Executive Board and Working Groups. As the report notes, "we need to ensure that the EOSC provides services that support innovative research and a competitive advantage for R&D in Europe 5 to 10 years from now and beyond that timespan."

Summary

The report lists several services (a "wish list") that were identified in discussions. These are highly diverse, from services for measuring impact of research, for machine translation, for assisting with ethical compliance, for regulation of access to data, and so on.

[5] https://doi.org/10.5281/zenodo.3693914

Implications for integration of the PID Graph

PIDs appear explicitly twice in the report: for assisting with computation of advanced metrics, and to identify different parts of an analysis or components of the research process. But their presence is implicit in other ways: for example, the proposed service "to identify all research output produced by one institution, researcher, network …" is a classic application of PIDs and the PID Graph. Integration of the PID Graph requires a clear recognition of the Graph as a component of infrastructure in its own right, on which services can be built, rather than just an enrichment of existing PID services.

Base document (2)

ESFRI cluster projects: Position papers on expectations and planned contributions to the EOSC (February 2020)[6]

Context

The ESFRI cluster projects were launched in 2019 to connect important communities to EOSC. The EOSCsecretariat project gathered stakeholder views on the implementation of EOSC, and this report presents five position papers from the cluster projects on their expectations, the added value of EOSC to the existing research infrastructure, issues to be addressed by the EOSC Executive Board and commitments to EOSC.

Summary

The common viewpoints (recommendations) emerging from the projects' viewpoints include:

- EOSC needs to aspire to offer a continuous, trusted working environment and networking opportunities to the research community.
- EOSC needs to provide a long-term open data archive with high performance storage and computing services, to enable sustainable use of data beyond the life span of individual data infrastructures.
- Collaborate with publishers to make open data a publication in its own right.

In addition, the ENVRI-FAIR project (in the environmental science domain) lists as one of its requirements that fundamental infrastructure components and metadata services (AAI, PID, provenance, workflow management, etc.) need to be integrated at sub-domain and RI levels, but also across the entire cluster.

Implications for integration of the PID Graph

Apart from the above-mentioned reference by ENVRI-FAIR, PIDs do not make an explicit appearance in the discussions. However, the requirements for trust and sustainability of data resources point to the need for PIDs and their supporting infrastructure, which itself must be sustainable. The PID Graph is not so clearly adumbrated as elsewhere, perhaps because of the already established and routine use of PIDs in these cluster projects' disciplines. There is a strong emphasis on access to, and reuse of data, and here we can envisage integration of the PID Graph as an enabler through linking related datasets, though it is clearly not all that is needed (the work of the FAIRsFAIR on metadata catalogue integration is also very relevant).

[6] https://doi.org/10.5281/zenodo.3675080

Base document (3)

Version 0.8 of the Strategic Research and Innovation Agenda (SRIA) for EOSC[7]

Context

The Strategic Research and Innovation Agenda is a roadmap for realising EOSC and will support the European Commission in developing the next work programme Horizon Europe. The current draft is version 0.8; the first full Version 1.0 of the SRIA is expected to be published by the EOSC Association in January 2021.

Summary

The SRIA is a complex and lengthy document (188 pages) that starts with "history and landscape of the digitisation of research in Europe" that led to EOSC. It surveys the policy context, guiding principles and the development of EOSC and its governance. Fourteen action areas are identified to help deploy the EOSC ecosystem, of which seven areas relate to the primarily technical challenges and prerequisites to implementing the EOSC ecosystem. These seven are:

- identifiers;
- metadata and ontologies;
- FAIR metrics and certification;
- authentication and authorisation infrastructure;
- user environments;
- resource provider environments;
- EOSC Interoperability Framework.

In each of these areas gaps and priorities are listed.

Implications for integration of the PID Graph

Unlike many other EOSC-related documents—with the exception of the PID policy dealt with below—the SRIA explicitly addresses PIDs as part of EOSC in the section on "identifiers" under the implementation challenges. What is more, it references the PID graph by name: one of the challenges is to "Provide standardised interfaces and protocols for exchanging information on PIDs to support the creation and use of a PID graph." PIDs are referenced in other contexts too. The EOSC-Core must assemble a number of basic services including PID services complying with the EOSC PID policy—specifically, services to generate, resolve and validate persistent identifiers. PID services are needed for FAIR for software and for the assessment framework for FAIR.

Some priorities for PIDs are listed, including a number to which FREYA has already contributed such as identifiers for resource types that are not yet established (on which FREYA had a whole Work Package and set of deliverables) and standardised interfaces for exchange of information to support a PID graph. There are some other challenges, too, including a "meta resolver" that can deal with any type of relevant identifier, and tools to support the certification of PID infrastructure against the EOSC PID Policy.

Here we see integration at several levels. The PID Graph is a conceptual part of EOSC, enabling "a rich, searchable resource for finding and contextualising resources". It is a concrete component of the infrastructure (that already exists, even if its relationship to EOSC Core is not yet completely clarified); yet there is still a need for further regularisation ("There is a need to standardise approaches across PID providers") and for "the uptake of tools built on this graph to become more widespread."

[7] https://www.eoscsecretariat.eu/news-opinion/eosc-strategic-research-and-innovation-agenda-version-08
https://www.eoscsecretariat.eu/sites/default/files/eosc-sria-v08.pdf

3.3 EOSC PID policy

Base document

"A Persistent Identifier (PID) policy for the European Open Science Cloud" (delivered to the EOSC Executive and Governance Boards Oct 2020)[8]

Context

The EOSC PID policy was a joint effort of the EOSC FAIR Working Group and Architecture Working Group and has now been formally published as an Independent Expert Report. Members of the FREYA project team were co-authors of the policy as part of their work on FREYA. The policy is written for senior decision makers within potential EOSC service and infrastructure providers, and will be of interest to all EOSC stakeholders. The expectation is that the policy should be maintained by the EOSC Legal Entity (Association) and reviewed by a dedicated group of experts and stakeholders on a regular basis.

Summary

The policy recognises that human use of PIDs and automated processing are both important. It recognises that PIDs may identify either digital objects or real-world entities, and it stresses interoperability of PID services between providers and across research infrastructures. It seeks to accommodate mature and established PID systems and practices, while not favouring one over another[9]. It also recognises the importance of GDPR in constraining the use of personal information, and the implications for PID systems.

The policy defines a number of components and roles with responsibilities in a PID infrastructure. The components are: PID Scheme; PID Service. The roles are: PID Authority; PID Service Provider; PID Manager; PID Owner; End User.

It identifies a need "for a generic, global PID resolution system across all PID systems and service providers. To enable this, PID Service Providers need to ensure their system supports the necessary API." (And this is precisely what a metaresolver like identifiers.org does.)

In terms of governance, responsibility is devolved on to PID Service Providers to provide community governance ensuring compliance with the policies; and implication is that the EU research community needs o be represented. However the governance structure should be embedded in global governance.

On long-term sustainability, the policy declares that "PID Service Providers should have a public and independently verifiable exit plan that assures continuity of their PIDs and PID Services should they cease to operate."

Implications for integration of the PID Graph

The EOSC PID Policy does not mention the PID Graph by name, in contrast with SRIA. However, there is recognition of the opportunities that the PID Graph offers: "The Policy should encourage new and innovative services and tools, which use and build on the PID Infrastructure" (section 2.9). The emphasis on interoperability is also important for enabling a rich and extensive PID Graph. In fact we could say that the policy is in part motivated by an awareness of the potential of the PID Graph and a desire to support it; the recognition of a diversity of entities identified by PIDs is also key here. Therefore the PID Policy can be seen as a highly important example of what was identified above as "enhancing uptake, capabilities or support for the PID Graph through EOSC".

[8] https://op.europa.eu/en/publication-detail/-/publication/35c5ca10-1417-11eb-b57e-01aa75ed71a1/language-en

[9] "A PID Service Provider should offer services that integrate well with European Research Infrastructures, but not at the exclusion of the broader research community." (section 7.1)

3.4 EOSC architecture

Base document

"EOSC-hub proposal for the EOSC Technical Architecture", slides from presentation by Giacinto Donvito, EOSC-hub Technology Coordinator, webinar 13 October 2020[10]

Context

The technical architecture for EOSC is complex and has been in a state of evolution for some time. The EOSC Pilot project proposed an elaborate service architecture, and EOSC-hub is developing a technical architecture and standards roadmap. The webinar given by Giacinto Donvito is a very recent and authoritative summary of the current thinking on the subject.

Summary

EOSC Pilot produced a service architecture with 47 classes of services split into five categories:

- front-end services;
- security and trust;
- Open Science, Data Management, Analytics;
- EOSC System Governance & Management;
- Compute and cloud platforms.

The EOSC-hub technical architecture details the different types and their relationships with other components of the architecture; the master diagram is reproduced in Figure 1

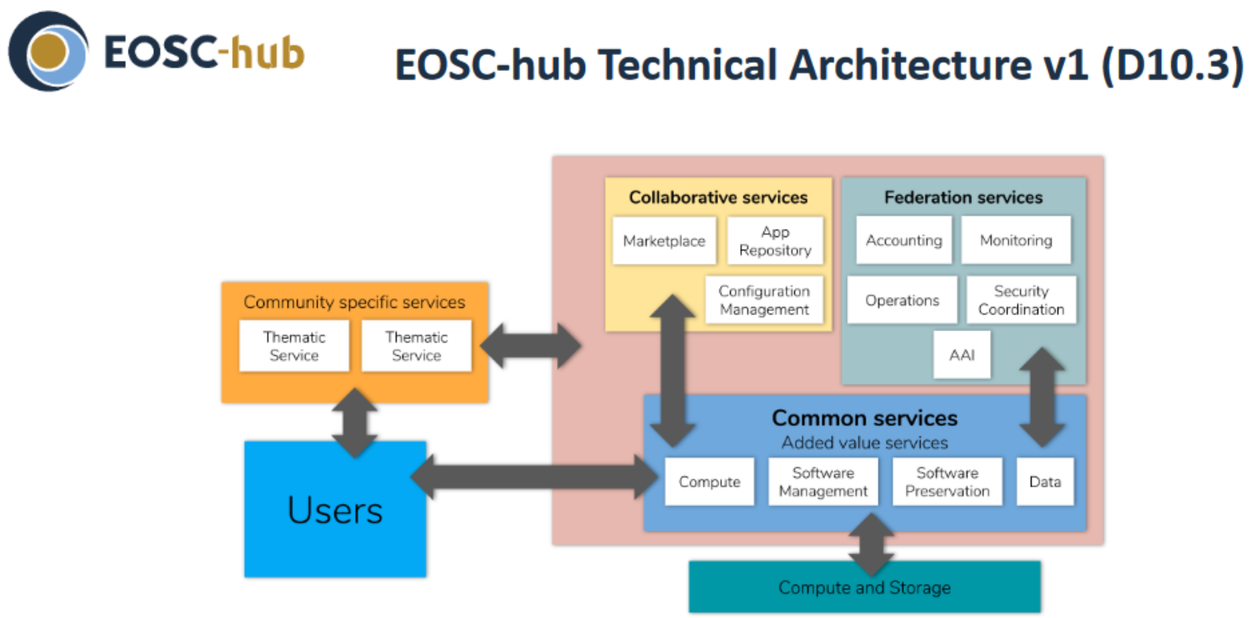

Figure 1 Overview of the EOSC-hub technical architecture

It is important that services can interoperate, and specifically be composable together, for example to create new scientific workflows, and the EOSC interoperability guidelines with standards and APIs encourage this. The EOSC service portfolio (which comprises both thematic and common services) sits

[10] https://www.eosc-hub.eu/sites/default/files/20201013%20-%20EOSC-hub%20Architecture%20-%20Main%20slide%20-%20v2.pdf

above the EOSC Federating Core which itself includes access-enabling services and federation services, as well as “shared resources”.

The approach taken is to define a reference architecture with a common vocabulary and listing the main “building blocks”. Each block will comprise a family of EOSC services and has its own technical specification. So for example there might be different monitoring tools but one building block for monitoring.

The architecture admits the possibility of other portals and marketplaces alongside the central EOSC Portal. The FREYA PID Services Registry was built to align with the EOSC marketplace.

Implications for integration of the PID Graph

PIDs are not prominent in the EOSC-hub technical architecture, and they do not feature in the list of “Technical specifications and interoperability guidelines”[11]. The FREYA project held a number of discussions with EOSC-hub to understand and clarify the relationship between FREYA’s work and the evolving EOSC, most recently in October 2020, and the following represents the consensus position at that time.

PID services are provided by independent organisations so cannot all be part of the EOSC Federating Core, which will be subject to EOSC governance. However PIDs are fundamental so should be part of the Minimum Viable EOSC (MVE), but this is still the subject of ongoing discussion. It is conceivable that some PID services might fall under EOSC’s “shared resources”, though that is not the main intent of the shared resources.

There is a particular need for PIDs to identify services registered in the EOSC Marketplace and in local marketplaces (communities, regional) so that connections can be made; and this should be part of the EOSC Core.

EOSC should provide guidelines for communities about assessing PIDs, and promote the guidelines to EOSC.

There are two directions to progress the integration of PIDs into EOSC:

- working to progress the common FREYA and EOSC-hub components through SRIA so that there is a roadmap for their future development—this would include FREYA’s PID Graph which is in fact already mentioned in SRIA;
- onboarding FREYA services into the Marketplace so that EOSC users can find PID services now (and indeed this is already being done).

[11] https://www.eosc-hub.eu/technical-documentation

4 Integration of FREYA outputs with EOSC

As explained in section 2, the meaning of "integration" of the PID Graph with EOSC is not clear-cut and can be understood at different levels. In order to accurately represent the integration that has been achieved or is required in future, it is helpful to adopt an existing framework. There is such a framework available, namely the European Interoperability Framework (EIF)[12]. This framework was established to give specific guidance on how to set up interoperable digital public services—it is not connected with EOSC, or indeed with scientific research at all, but it embodies a thorough analysis of the challenges of delivery of services in a diverse environment with many actors and a fast-moving digital world.

The EIF defines interoperability as:

> *… the ability of organisations to interact towards mutually beneficial goals, involving the sharing of information and knowledge between these organisations, through the business processes they support, by means of the exchange of data between their ICT systems.*

This echoes the vision of EOSC as a federated infrastructure of interacting components providing seamless access to data and services. In this context, interoperability as a goal corresponds closely to the "integration" (or "embedding") in EOSC to which FREYA aspires.

A key component of the EIF is its interoperability model. For the purposes of this analysis, the cross-cutting component "integrated public service governance" can be set aside, as can the background layer "interoperability governance". What is important is a representation of the dimensions or layers of interoperability; its governance is a different concern, though one with EOSC will of course be engaged. The four layers of interoperability are (illustrated in Figure 2):

- **Legal:** ensuring that organisations operating under different legal frameworks, policies and strategies are able to work together.
- **Organisational:** alignment of business processes, responsibilities and expectations to achieve commonly agreed and mutually beneficial goals.
- **Semantic**: ensures that the precise format and meaning of exchanged data and information is preserved and understood throughout exchanges between parties.
- **Technical:** covers the applications and infrastructures linking systems and services.

These four layers have meaning in the context of EOSC, and provide a structure for analysing the integration or embedding of technologies, processes, practices and governance that exist independently of EOSC.

[12] https://joinup.ec.europa.eu/collection/nifo-national-interoperability-framework-observatory/eif-european-interoperability-framework-0

Figure 2 The layers of interoperability in the European Interoperability Framework[13]

The FREYA team has produced a graphic of the results of FREYA related to each of the three pillars of FREYA, shown in Figure 3. This graphic illustrates the coverage, influence and diversity of the results. From this, it is possible to extract those results related to the PID Graph that could or should be integrated with EOSC. The criterion for inclusion is that they are sustained services or applications, rather than static outputs that contributed during the lifetime of FREYA but will not be continued; and also that they are not organisationally completely independent of EOSC (as the RDA groups are). The selected outputs are:

- technical bases of the PID Graph such the GraphQL API;
- ancillary services such as common DOI search ("DataCite Commons");
- the PID Services Registry;
- disciplinary applications of the PID Graph, in as much as they can be made available as services for wider communities beyond the use only of the FREYA partner that developed them;

In addition, there are other outputs of FREYA that are sufficiently important for EOSC that, though they do not meet the above criteria they should be considered from the perspective of integration:

- EOSC PID policy work (to which FREYA contributed in the lifetime of the project);
- *pidforum.org* which while independent of EOSC can surely play a role in bringing together stakeholders and reflect a global perspective on PIDs of which EOSC must be aware.

Finally, another key output of FREYA, the PID Commons has its own deliverable D6.5 and is not discussed here. The relationship between a global PID Commons and EOSC will necessarily be complex and it is not possible at this stage to talk of "integration"—though if and when it takes place it will surely span the four levels of interoperability.

[13] Source of figure: https://joinup.ec.europa.eu/collection/nifo-national-interoperability-framework-observatory/3-interoperability-layers

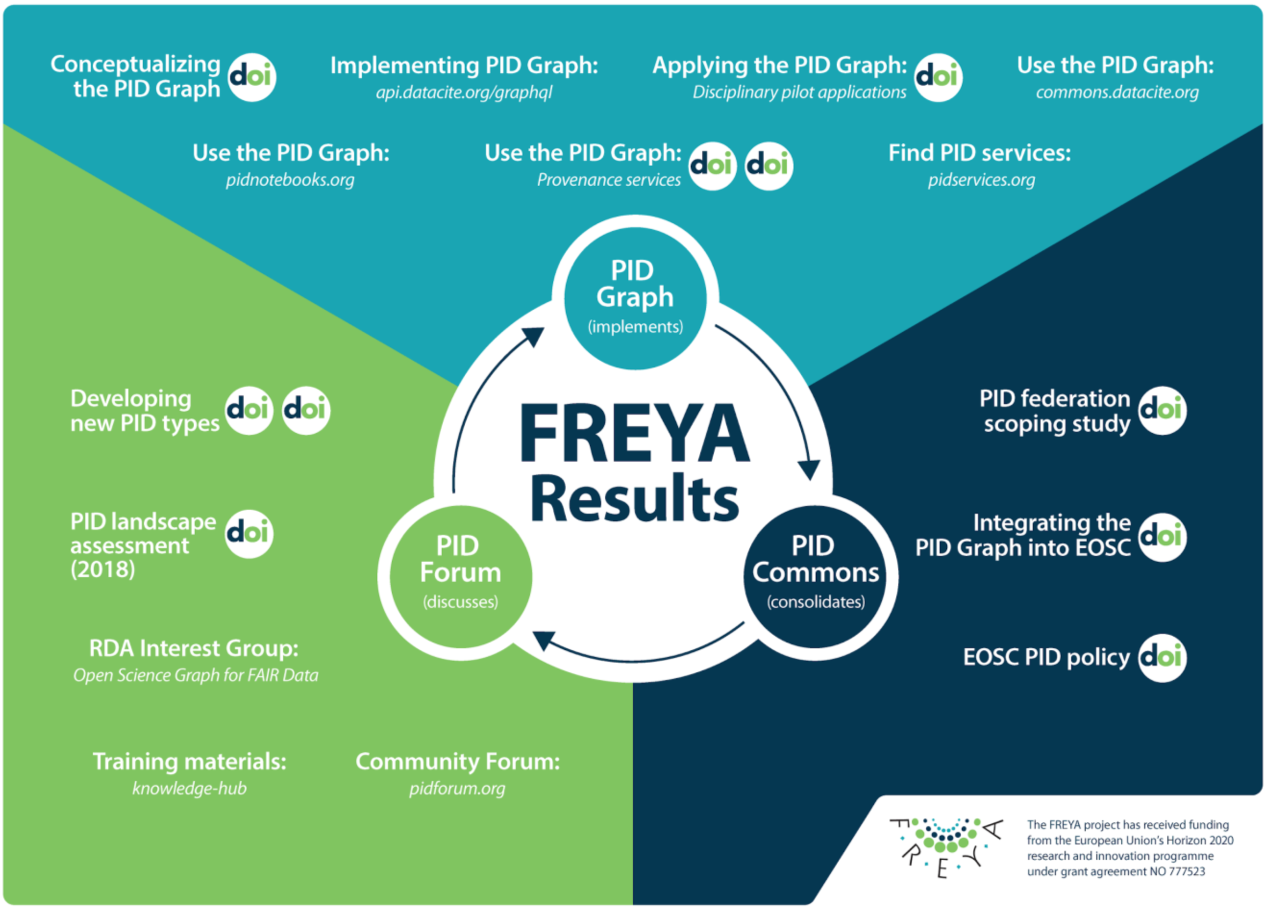

Figure 3 A depiction of the major results of FREYA

Considering then the six selected outputs of FREYA, the status of integration with EOSC can be represented as shown in Table 1.

FREYA output	Interoperability layer(s)	Status of integration
Technical bases of PID Graph	Semantic Technical	The services will be registered in the EOSC Marketplace. However, the PID Graph also plays a more fundamental role than simply a service available for EOSC end-users to develop their own applications. As seen in the discussion of SRIA (section 3.2), the PID Graph is seen as a valuable component of EOSC in its own right, and there may have a place in the EOSC Core. The discussion of the EOSC technical architecture (section 3.4) summarises the state at the end of the FREYA project.
Ancillary services	Semantic Technical	The services will be registered in the EOSC Marketplace.
PID Services Registry	Organisational	Route for integration into EOSC and relationship to EOSC catalogue to be decided.

Disciplinary applications	Organisational	Some applications developed or enhanced in FREYA are available through the EOSC Marketplace.
PID policy	Legal Organisational	The policy supports the vision and future of the PID Graph, as discussed in section 3.3.
pidforum.org	Organisational	Independent operation, but potential contributions to EOSC stakeholder communication should be pursued.

Table 1 Status of integration of FREYA outputs with EOSC

5 Engagement with EOSC cluster projects

Right from the start, FREYA has engaged with EOSC through many workshop, conferences, webinars and private meetings; these are summarised in the annual reports on engagement activities. In line with the task description, however, the project has also engaged with the EOSC cluster projects to foster PID service adoption and to understand their perspectives and uses of PIDs as part of their established infrastructures.

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?

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.

5.1 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.

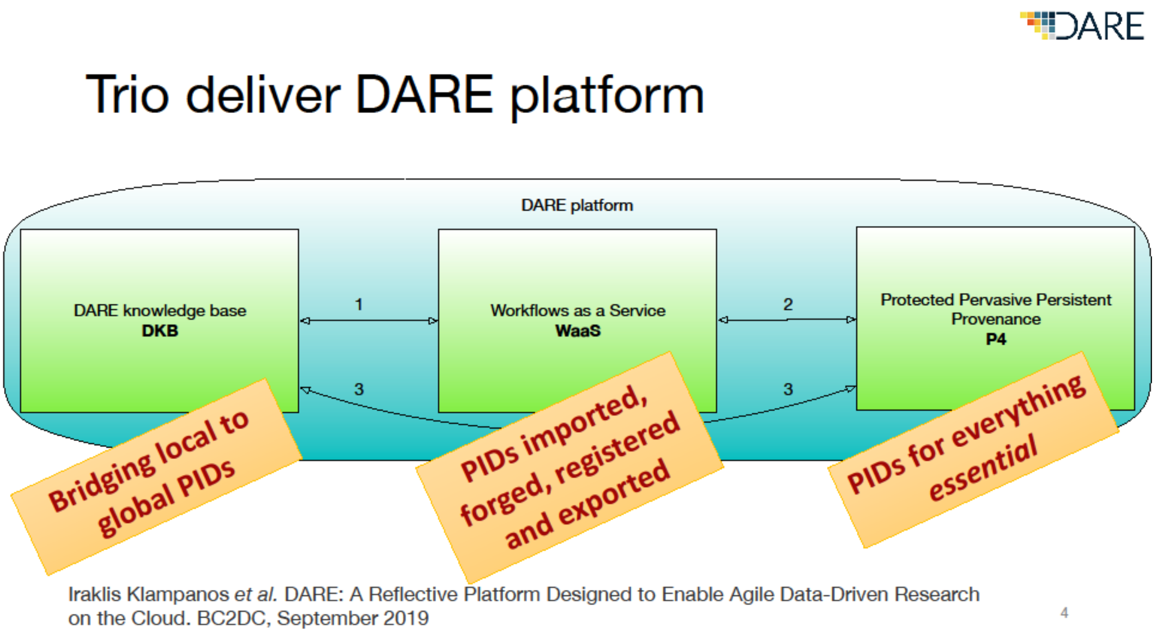

Figure 4 Depiction of the DARE platform

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.

5.2 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.

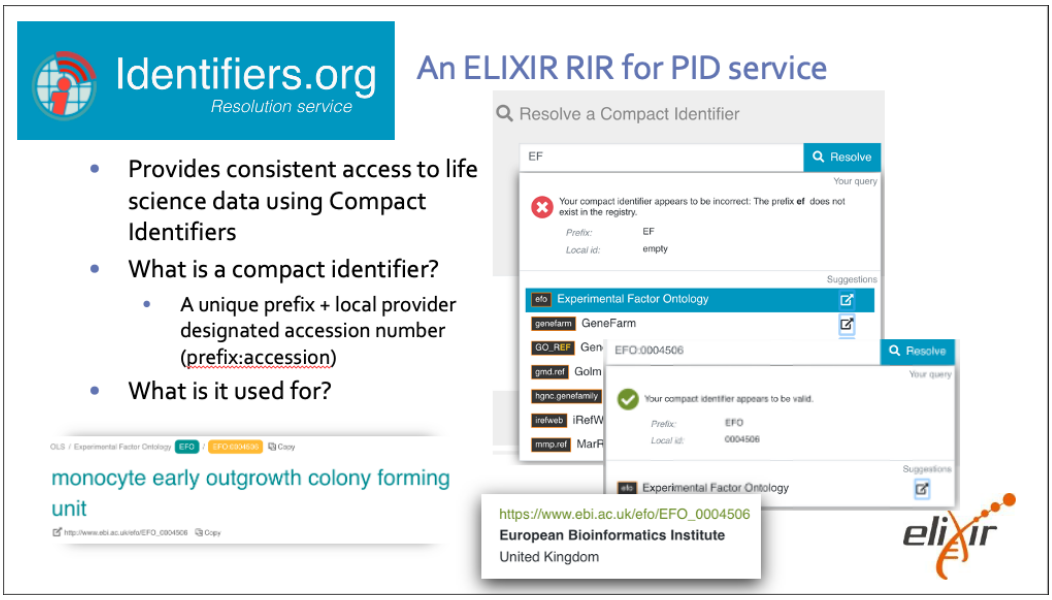

Figure 5 Identifiers.org in context

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.

5.3 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.

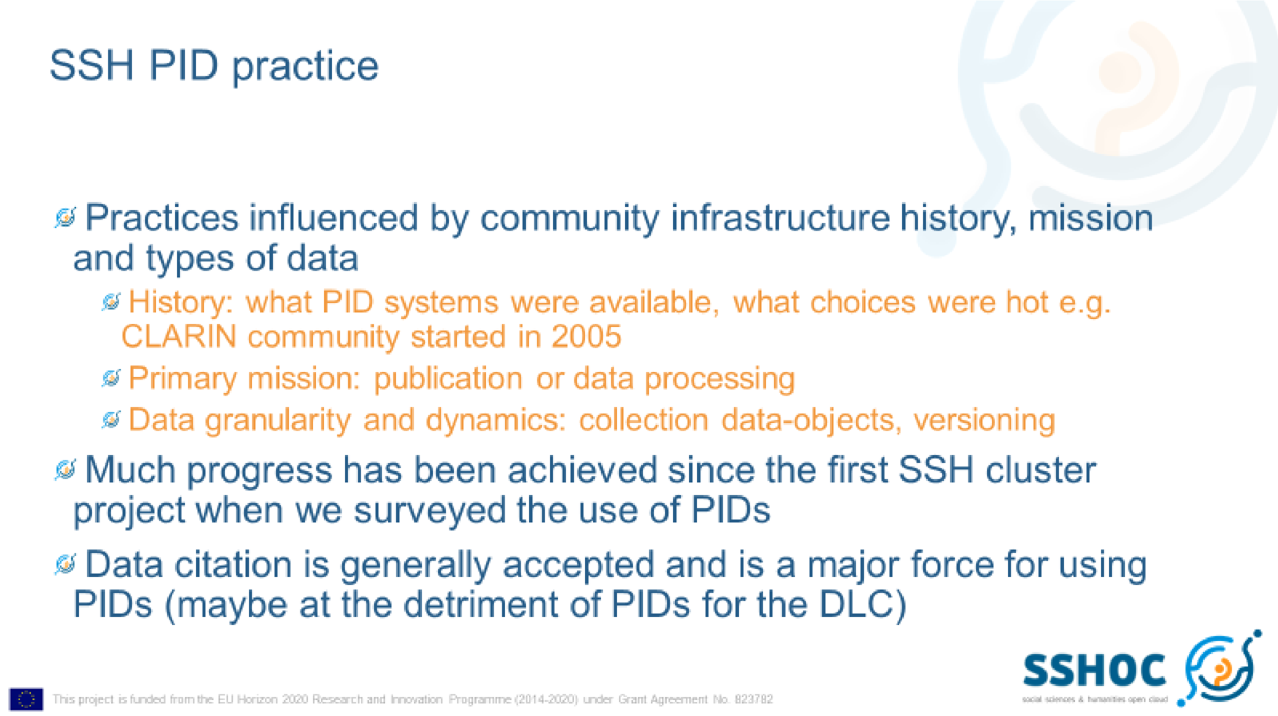

Figure 6 Overview of PID practice in the social sciences and humanities

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.

5.4 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.

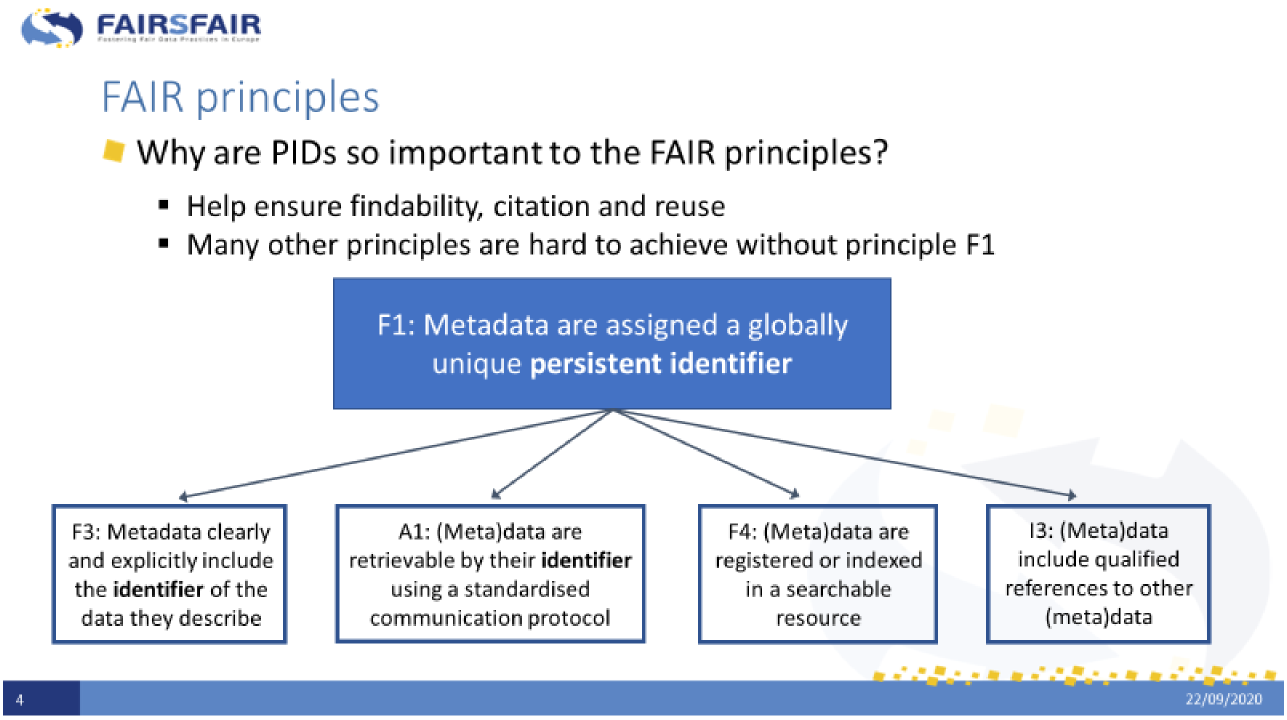

Figure 7 PIDs in the FAIR principles (FAIRsFAIR project)

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.

5.5 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).

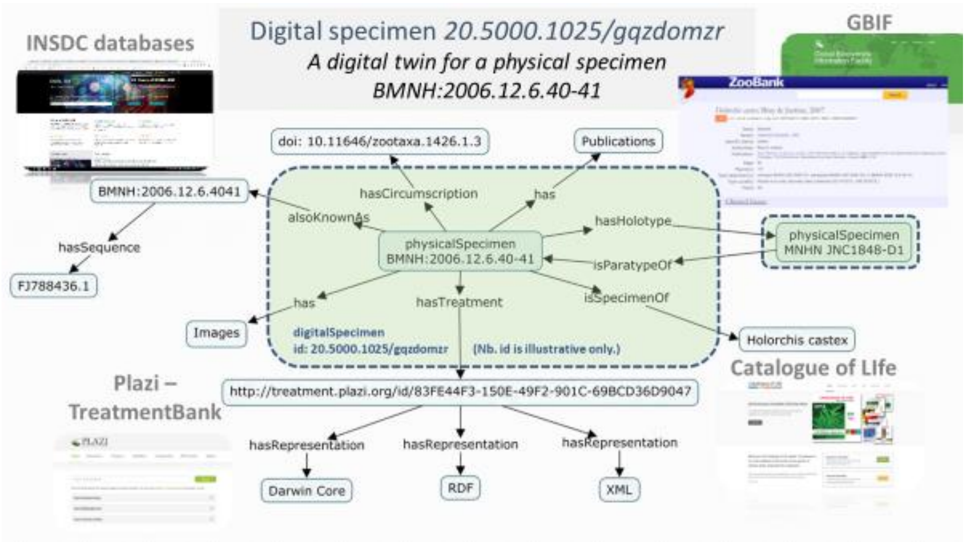

Figure 8 Digital representation of natural sciences specimens

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.

5.6 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.

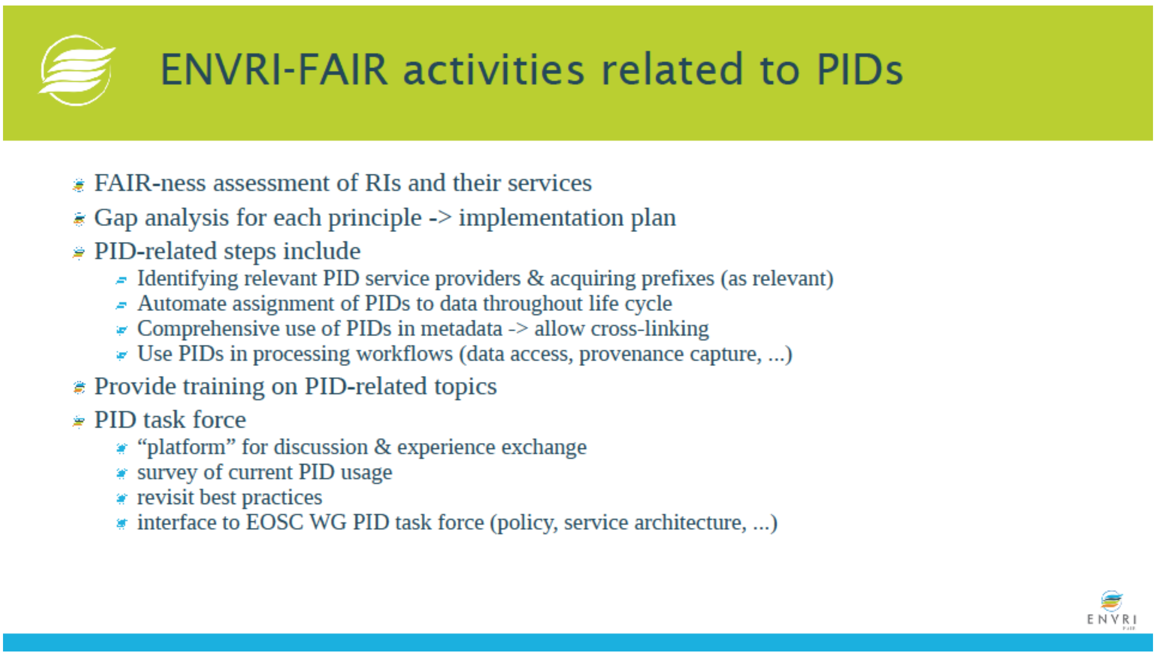

Figure 9 ENVRI-FAIR activities related to PIDs

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.

5.7 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 data discovery and PID use. Both 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).

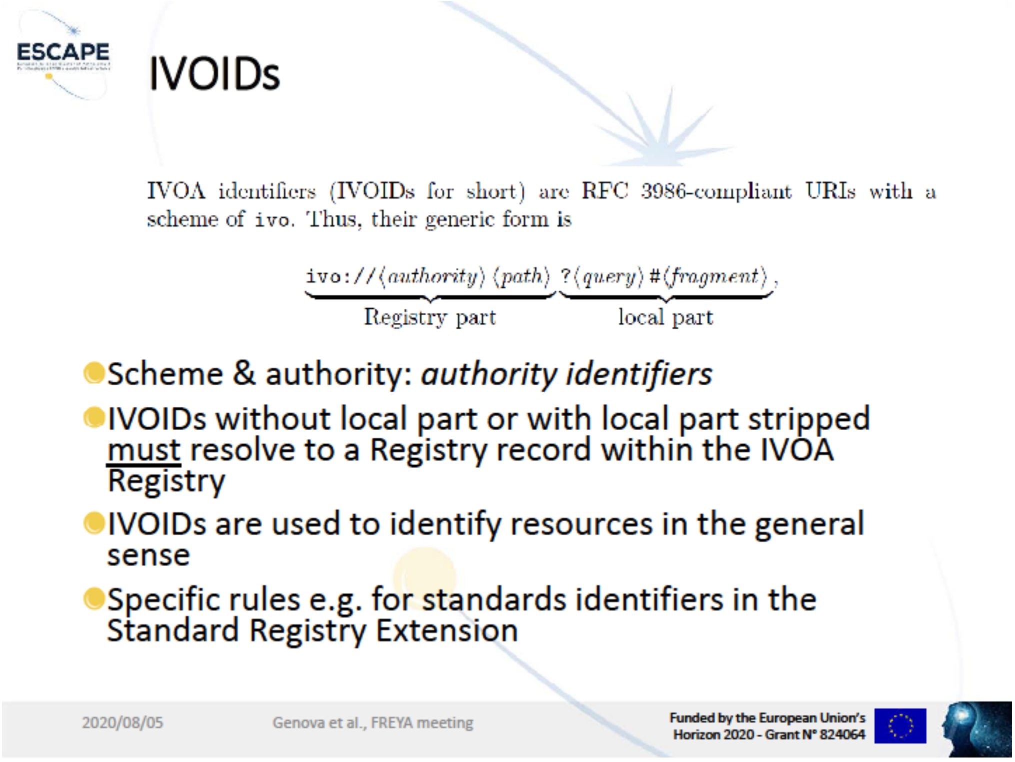

Figure 10 The basics of IVOIDs (ESCAPE project)

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.

6 Conclusions and next steps

The reconciliation of the dual motivations of FREYA—arising from the vision of the PID Graph and from the project's role in contributing to EOSC—have not been easy to reconcile, especially given the continuous evolution of EOSC which is still ongoing at the end of the FREYA project. Expecting a simple kind of "integration" of FREYA results into EOSC was not realistic, but nonetheless, by carefully analysing what levels of integration might exist, it has been possible to define the roles that key FREYA outputs play in EOSC, and indeed how EOSC can support the advancement of FREYA's own vision of the PID Graph.

Obviously, the job of integration is not finished. This deliverable will end with a handful of concise recommendations for the next steps in achieving integration of the PID Graph with EOSC, drawn from the discussions in the foregoing sections.

1. Ensure that the PID Graph appears in appropriate EOSC documents as a motivating vision (as it does in SRIA).
2. Use SRIA to set the roadmap for development of the PID Graph within the EOSC context.
3. Follow up the need expressed in the EOSC PID Policy for a generic, global PID resolution system across all PID systems and service providers.
4. Establish a clear relationship between the FREYA PID Services Registry and the EOSC Portal.
5. Take advantage of the success and wide reach of *pidforum.org* to engage with PID-interested stakeholders globally.