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

D4.7 Using the PID Graph: Community Workflows and Discoverability Services

Deliverable type	Report
Dissemination level	Public
Due date	30 November 2020
Authors	Artemis Lavasa (CERN, orcid.org/0000-0001-5633-2459) Stephanie van de Sandt (CERN, orcid.org/0000-0002-9576-1974) Pamfilos Fokianos (CERN, orcid.org/0000-0003-0618-1722) Christine Ferguson (EMBL-EBI, orcid.org/0000-0002-9317-6819) Henning Hermjakob (EMBL-EBI, orcid.org/0000-0001-8479-0262) Chris Baars (DANS, orcid.org/0000-0002-5228-1970) Ricarda Braukmann (DANS, orcid.org/0000-0001-6383-7148) Tina Dohna (PANGAEA, orcid.org/0000-0002-5948-0980) Uwe Schindler (PANGAEA, orcid.org/0000-0002-1900-4162) Frances Madden (British Library, orcid.org/0000-0002-5432-6116) Vasily Bunakov (STFC, orcid.org/0000-0003-3467-5690) Martin Fenner (DataCite, orcid.org/0000-0003-1419-2405)

Abstract This deliverable reports on services and pilots developed by the disciplinary partners in FREYA as a way of demonstrating how the PID Graph concept can be integrated into research and community workflows. This is the final deliverable of Work Package 4 and as such it presents the progress and final state of the disciplinary pilot applications, and summarizes lessons learned and potential future plans for PID Graph work.

Status Submitted to EC 27 November 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.

Executive summary

Throughout the lifetime of this project, the disciplinary partners in FREYA, namely the British Library, CERN, DANS, EMBL-EBI, PANGAEA and STFC, have been establishing and enhancing services which center around the PID Graph, a network of interconnected systems via persistent identifiers (PIDs), for the benefit of their individual research communities. This report is the final deliverable of Work Package 4, which is concerned with integrating the vision of the PID Graph into the practices of disciplinary communities. It presents the progress and final state of the disciplinary pilot applications since the beginning of the project.

Each disciplinary community has very different priorities, needs and degrees of adaptability in terms of embracing new PID-related developments in systems or services they regularly use as part of their workflows. This is evident through the work each partner chose to carry out, as well as the variety of technical issues brought up in the final chapter of this report. This created numerous challenges during the development of PID services due to community-specific behaviors which complicate one-size-fits-all solutions across disciplines. Each community faces different challenges and has distinct approaches and vision for PID Graph development for the future, nonetheless there were also some observations that were applicable to more than one community - for example, more and better support for users when it comes to using PIDs in their workflows, and increased outreach from the side of the service providers.

Moreover, in a separate strand of work, a number of Jupyter notebooks were developed in the last phase of the project as an additional method of enabling PID discoverability, which are highly adaptable and relevant in the current technological environment, and easily related to the needs of many communities.

The overarching goals of this Work Package are concerned with using the pilot applications to demonstrate how the PID Graph can be implemented in disciplinary services, and how communities can discover and exploit the content of the PID Graph. The tools and practices developed throughout the lifetime of the project will continue to be exploited by the FREYA disciplinary partners to further build on their services for the benefit of their communities.

Contents

1 Introduction

The objective of Work Package 4 (WP4) is to integrate disciplinary and EOSC contexts with the PID Graph. It builds on work from WP2 "PID Core Services" and WP3 "New PID Types" to demonstrate how disciplinary demonstrators in the FREYA project can access and use information from the PID Graph to benefit their individual communities. The work in this WP is meant to focus on exploitation and enrichment of the PID Graph by using the tools and services developed in WP2, and through taking over and extending the work introduced by WP3 about surveying the current PID landscape and the development of demonstrators for new/emerging PID types.

Previous deliverables in this Work Package addressed the following:

- D4.1[1] - Introduction of the disciplinary pilot applications and laying the groundwork for building the disciplinary PID graphs. Enhancement of person-article-data linking, further establishing software citation and publication workflows, and integrating mature PID types into the different disciplinary systems.
- D4.2[2] - Work on resource and metadata provenance within the disciplinary pilot applications in FREYA.
- D4.3[3] - Development of specific pilot projects to facilitate advancement of the existing PID Graph functionality in community services.
- D4.4[4] - The first part of the task of integrating emerging PID types into disciplinary contexts, focusing on integrations of persistent identifiers for organizations.
- D4.5[5] - Integration of the PID Graph with EOSC by means of assessing the EOSC landscape, determining how FREYA outputs relate to the needs of the EOSC and understanding the needs of the EOSC by engaging with EOSC projects.
- D4.6 - The second part of the task of integrating emerging PID types into disciplinary contexts, focusing on a variety of PID integrations, from organization IDs, to grant and funder IDs, project IDs, etc.

D4.7 (the present document) summarizes work done within WP4 throughout the lifetime of the project and introduces final developments by the FREYA disciplinary partners. More specifically, it addresses task T4.3.2 "Refinement and enhancement of community-specific workflows with PID Graph functionality" which is about implementing workflows to better exploit PIDs and to further incentivize the usage of PIDs within research workflows, and task T4.3.3 "Advanced disciplinary PID Graph discovery" regarding enabling end-users of specific WP4 pilot applications to discover the content of the PID Graph.

Pilot applications in this context refer to concrete developments that exploit a PID graph in some way to address relevant user stories or needs in a particular discipline. **The** PID Graph refers to the PID Graph core infrastructure work developed by the project on the whole, and specifically the services developed by WP2, most notably the DataCite GraphQL and, by extension, the Jupyter notebooks. The entire concept and work on the PID Graph is presented in a separate report[6]. **A** PID graph, when referring to the work of the pilot applications, can be any "local" graph of interconnected PIDs that supports a particular service; in those cases it refers to a subset of PIDs, a disciplinary subgraph that may make use of the PID Graph tools developed by WP2. It is important to note that WP4 does not only pertain to developing software and services, but also defines workflows and how communities use (local) PID graphs, and how they could extend them.

[1] D4.1: https://doi.org/10.5281/zenodo.2414838
[2] D4.2: https://doi.org/10.5281/zenodo.3249832
[3] D4.3: https://doi.org/10.5281/zenodo.4066841
[4] D4.4: https://doi.org/10.5281/zenodo.3606059
[5] Links are unavailable for D4.5 and D4.6 as they were published at the same time or just before this report.
[6] The FREYA PID Graph: https://doi.org/10.5281/zenodo.4028382

As this is the final deliverable of WP4, Chapter 2 presents the final state of the disciplinary pilot applications in FREYA and previous work that has been carried out by the FREYA disciplinary partners, namely the British Library (Humanities and Social Sciences), CERN (High-Energy Physics), DANS (Social Sciences), EMBL-EBI (Life Sciences), PANGAEA (Earth and Environmental Sciences) and STFC (Facilities-based Science), to advance these pilots. This deliverable presents development work carried out by disciplinary partners for their pilot applications to increase and incentivize the usage of PIDs, and to enable discovery of PID Graph content; this is work that either enhances an already-existing (prior to the FREYA project) service/tool or work that introduces an entirely new service/tool.

Early in the project, FREYA partners gathered a number of PID-related user stories which reflected the needs of their individual community and required PID Graph functionality. During the final year of the project, the user stories were the basis of another piece of work that is reported in Chapter 3 of this deliverable, which is the development of a number of computational notebooks. These Jupyter notebooks[7] were developed as part of a subcontract. They are based on specific use cases that the project thought could be addressed given the available time frame of the project. The notebooks show how the PID Graph can be queried through the GraphQL API[8], and can be used as a method to incentivize communities to interact with the PID Graph. This work demonstrates that FREYA partners could potentially be "suppliers" of GraphQL APIs and not just providers. The notebooks can be adapted to address similar use cases beyond the ones developed by FREYA to facilitate discovery of more "complex" PID Graph information (e.g. interconnections between PIDs, reuse tracking, etc.). This development was valuable because the generic nature of the use cases can be related to a wide range of stakeholders and disciplines, not only those represented by FREYA's own disciplinary partners.

The final chapter of this deliverable, Chapter 4, discusses lessons learned by each partner and general plans for the future regarding the PID Graph and its value for each disciplinary community represented in FREYA.

[7] As stated in the Project Jupyter website (https://jupyter.org/): "The Jupyter Notebook is an open-source web application that allows you to create and share documents that contain live code, equations, visualizations and narrative text".

[8] DataCite GraphQL API: https://doi.org/10.5438/yfck-mv39

2 Using the PID Graph: disciplinary pilot applications

2.1 British Library

Pilot applications

The British Library's pilot applications within FREYA include data.bl.uk, the British Library's collection of derived datasets and the UK's index of theses, EThOS[9]. The data.bl.uk datasets collection is now hosted on the British Library's pilot Shared Research Repository[10] and EThOS will likely be migrated to that platform soon. We also undertook a small amount of development work in the general library catalogue, Explore[11].

Shared Research Repository
The British Library's Shared Research Repository is a multi-tenanted repository for cultural heritage organizations in the UK. At present, six museums and other heritage organizations avail of the repository service, administered by the British Library, which provides individual research repositories and a shared searchable layer. These repositories are designed to provide a unified location for the research conducted within these organizations to be found. The repository service, based on Samvera Hyku[12], was launched as a beta service in October 2019 and will go into full service in January 2021. The content of these shared repositories include outputs from the research conducted by the staff of these cultural heritage organizations, the students who undertake collaborative doctoral studies with these organizations, and university and doctoral student placements.

EThOS
The UK's index of doctoral theses, EThOS is administered by the British Library. It is due to be migrated to a new platform in the coming years, which will likely be the same platform as the research repository. Therefore these developments will integrate with that replatforming when it takes place in 2021 onwards.

Explore
The British Library's main catalogue, Explore, is based on Primo. For D4.3 we undertook a small amount of work using a PID graph to improve the discovery of journal articles within that service.

Technical work

Shared Research Repository/EThOS
Within the research repository, technical work included increasing PID graph functionality such as supporting emerging PID types and improving support for related identifiers by making those links actionable on the front-end of the repository. At its launch in 2019 it supported ORCID iDs and ISNIs for creators and contributors of repository records. Now, the repository also includes support for ROR IDs, Wikidata IDs and GRID IDs in the organizational creator and contributor fields, (see Figure 1).

The Crossref Funder Registry is included as a lookup for the "funder" field which now includes a link to that funder's ROR and ISNI where available, as illustrated in Figure 2. Multiple funders and project references can now be included associated with a specific funder. For the Doctoral Thesis template, in light of the anticipated EThOS migration, identifiers were included in the table which provides a controlled list of the UK universities which contribute to EThOS. This work was reported in D4.4 and D4.6 and completed in August 2020.

[9] EThOS: https://ethos.bl.uk/
[10] The British Library's pilot Shared Research Repository: https://iro.bl.uk/
[11] Explore: http://explore.bl.uk/
[12] Samvera Hyku: https://hyku.samvera.org/

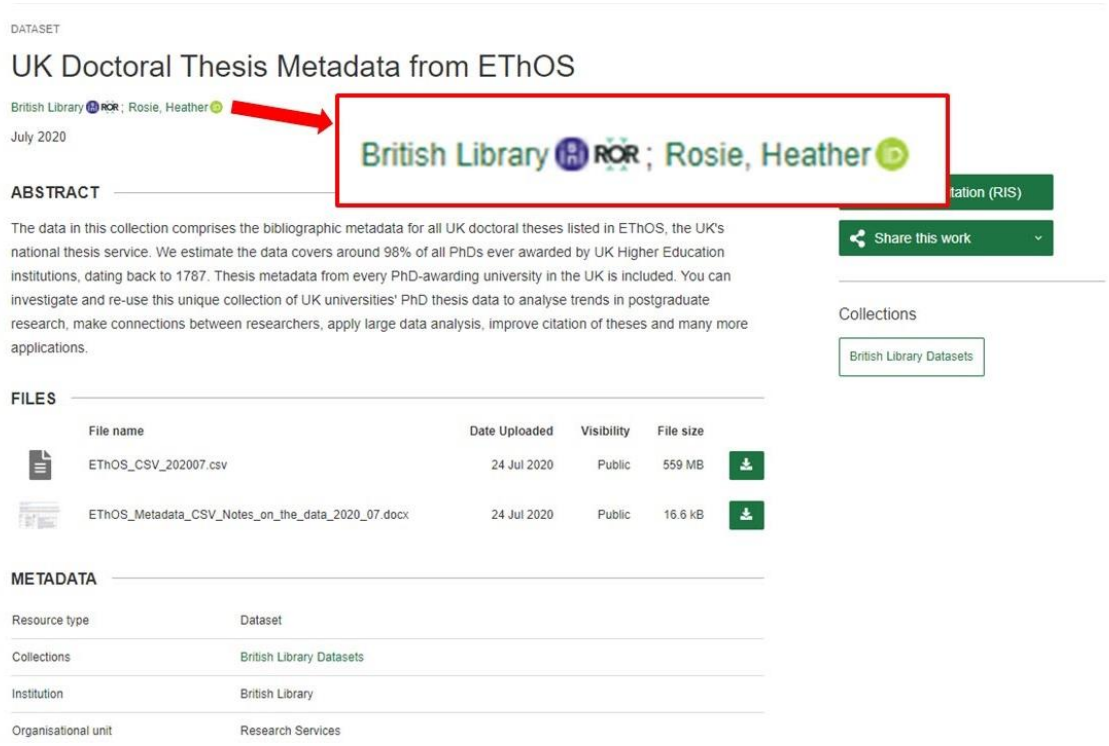


Figure 1: A screenshot showing the support for ROR in the Creator field of the British Library's Shared Research Repository

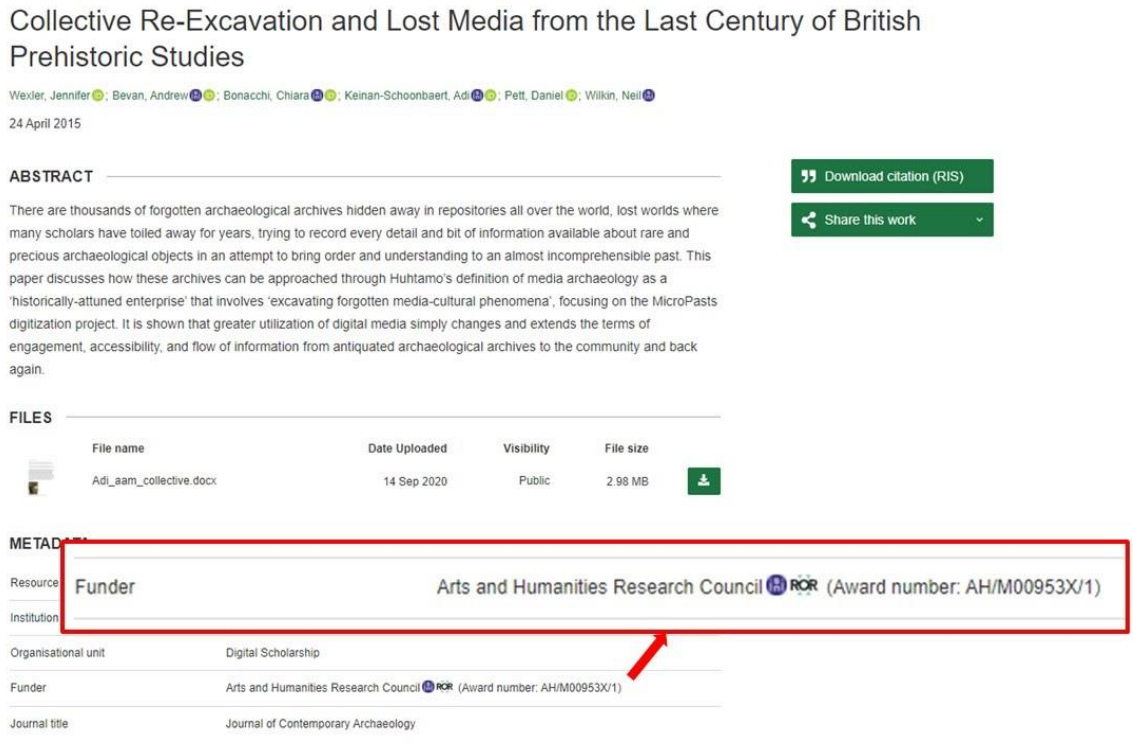


Figure 2: A screenshot demonstrating the revised support for PIDs in the Funder field of the British Library's Shared Research Repository

Explore

This piece of development improves the accessibility of journal articles in our catalogue by addressing an issue where there are multiple catalogue records for the same resource but they all provide separate access routes to the resource, some of which require a user to be on site. This solution uses the DOI of the article to populate the "I want this" tab information with all of the access options available for the resource. This work was reported in D4.3 and was completed in November 2020.

Jupyter notebooks

The British Library decided to take the Jupyter notebooks created within FREYA which query the GraphQL API (Chapter 3.2) and apply them to their own outputs. One of the notebooks yielded results to explore the publication of datasets across the British Library over time and is available on the dedicated British Library Labs GitHub repository[13].

PID graph

The British Library's PID graph has been extended through the inclusion of ROR IDs, Crossref Funder Registry DOIs, GRID and Wikidata identifiers, and by making related identifiers such as ARK identifiers, which are assigned to digitized and born-digital collection items, actionable. This allows users to navigate through the PID graph to find other entities at an additional remove from the dataset. Figure 3 shows where the graph has been extended.

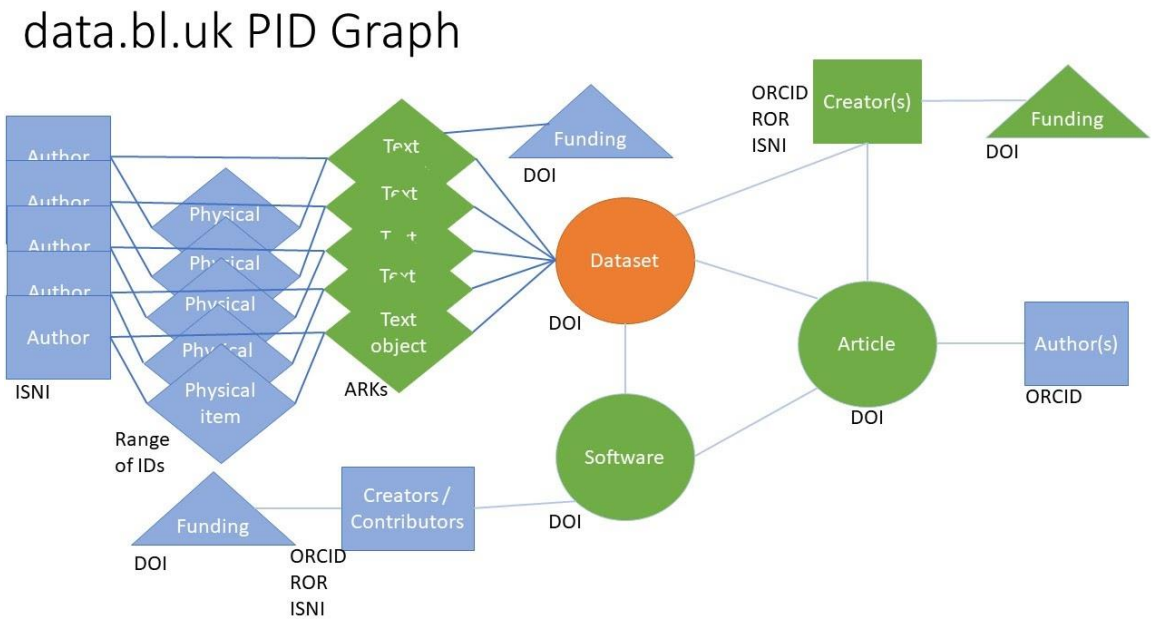


Figure 3: This diagram indicates how the graph for a data.bl.uk dataset has been extended through FREYA. Items in green indicate where related identifiers or other entities can now be included as actionable entities in the Shared Repository. Items in blue can be discovered through navigating through those in green.

Throughout the development of the Shared Research Repository we have supported the functionality of PIDs and made them discoverable to users. By including ISNI, ORCID and ROR as logos adjacent to the creator and contributor name, we are encouraging their usage and awareness to the community, where they are not always used as standard. The team creating the majority of the datasets for the data.bl.uk collection, have been aware of the benefits of DOIs for their materials for a long time but as the Shared Research Repository has been promoted to the wider library staff, the benefits of DOIs have been communicated and realized.

[13] British Library Labs GitHub repository: https://github.com/BL-Labs/Jupyter-notebooks-projects-using-BL-Sources/

As the Shared Research Repository has several tenants beyond the British Library, including the British Museum, Tate and National Museums Scotland, the possibilities and capabilities of PIDs have been illustrated there. By having a DOI lookup for both Crossref and DataCite DOIs, as core functionality of the repository since its launch, the administrators of the repositories in those areas can see the benefits of using shared infrastructure. In addition, the functionality of emerging PIDs will be available to them too when they are adding records to their repositories. These developments will also be made available to any Samvera Hyku user as this functionality will be merged into the Hyku codebase as an output of the Advancing Hyku project[14].

Through the work undertaken in D4.3, the British Library is allowing users to navigate more clearly to electronic resources with DOIs, such as journal articles accessed through ARK identifiers via the item's DOI.

Lessons learned

1. From the humanities perspective, especially within a national library, it is necessary to provide identifiers for a broad range of content, across a large timeframe. This has meant that there is a need to support identifiers appropriate to historical as well as current content. From this perspective, identifiers such as ISNI are important as they have that capability which ORCID or ROR do not.
2. The British Library felt it was important that a number of different types of organizational identifier be supported within the repository to ensure as broad a swathe of coverage as possible and to future-proof the repository. In turn this presented a readability challenge for end users who would be presented with large numbers of hyperlinked icons with which they were not aware. To address this, it was decided to limit the number of identifiers displayed on the front-end of the repository.
3. In the future, the British Library would like to improve the integration between ARK identifiers which are used for digitized and born-digital collection items, and the Library's systems. These ARK identifiers are not externally resolvable, and while some scoping work for making them externally resolvable was undertaken within FREYA as part of the work in Work Package 3, it was not possible to make that change within the timeframe of the project.

2.2 CERN

Pilot applications

As part of CERN's contribution to WP4, various existing CERN services that use PIDs were enhanced with more features and functionalities, namely CERN Analysis Preservation, the CERN Open Data portal and INSPIRE. CERN Analysis Preservation (CAP)[15] is a restricted-access service currently in beta phase. It is a tool for researchers at CERN to describe and preserve all the components of a physics analysis (e.g. data, software and computing environment) to ensure that they are reusable and understandable in the future. The CERN Open Data portal[16] is an open-access repository which currently holds over two petabytes of particle physics data and accompanying code/software, and documentation produced through the research performed at CERN. INSPIRE[17] is the core High-Energy Physics (HEP) information system aggregating content from multiple sources.

While CAP, INSPIRE, and the CERN Open Data portal have acted as CERN's main pilot applications in FREYA, relevant work on other CERN-based services, notably the multidisciplinary open science repository

[14] Advancing Hyku project: https://advancinghyku.io/
[15] CERN Analysis Preservation: https://analysispreservation.cern.ch
[16] CERN Open Data portal: http://opendata.cern.ch/
[17] INSPIRE: https://inspirehep.net/

Zenodo[18], has also in some cases been presented in WP4 deliverables which helps with providing an overview of PID-related work at CERN more generally.

Technical work

CERN's technical work during the lifetime of FREYA has focused on better linking of resources through PIDs (e.g. linking data with publications and software) and expanding that to include new PID types, on making data and software a first-class citizen by strengthening efforts to allow users to get DOIs for them through various services for better credit attribution and impact tracking, and further advancing metadata enrichment to further support reproducibility of research results.

Incorporating new PID types

ROR IDs have been integrated into INSPIRE, CERN Open Data and CAP as pilots so far and they will also be supported by Zenodo soon. On INSPIRE, GRID was already used in the past and currently both identifiers are supported. This work was presented in D4.4 and since then the ROR integration on INSPIRE has been moved to the production system and at the same time INSPIRE migrated to a new version which also upgraded the look of the organization records. For CAP, support for organization identifiers was introduced as a completely new feature and it was decided to enable fetching for ROR IDs only. For CERN Open Data, a Minimum Viable Product (MVP) for ROR IDs was introduced in October 2020 for a selected number of records in the production system (see more in D4.6).

Capturing metadata from various external services through PIDs and adding it to CAP records has been one of the main tasks for CAP. CAP can be used as an information hub for physics analyses as it brings together information from various tools and databases, both open as well as restricted. Other than the Zenodo and ORCID integrations that have already been presented in D4.3, information can now be fetched from the ROR database as well to populate affiliation information in CAP (Figure 4).

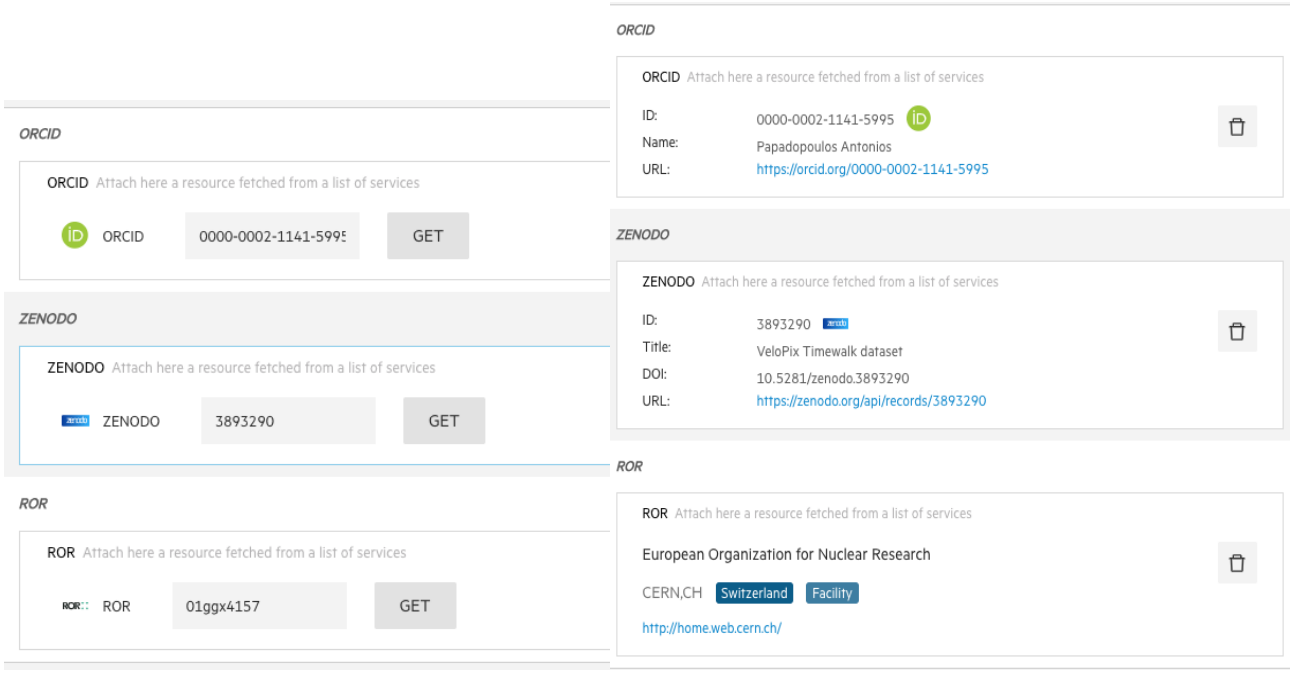


Figure 4: Integrations of external PID services in CERN Analysis preservation. The user can fetch information from ORCID, Zenodo and ROR and populate their analysis record. The left image is how the form looks when a user asks for information using a PID or a record ID; the image on the right is the information fetched from the databases as it appears on a record.

[18] Zenodo: https://zenodo.org/

While we cannot mint DOIs for analyses in CAP because at that stage they are confidential/restricted, within CAP, each analysis record is assigned a unique ID (shown at the top of Figure 5 below) which can take a more "meaningful" display format once the analysis is no longer in a draft state (e.g. "CAP.CMS.XXXX.YYYY" to identify an analysis by the CMS collaboration). Future plans for CAP include enabling users to prepare these complex research outputs for public releases and pushing information to public-facing publishing platforms directly from CAP, which can then "transform" the internal CAP analysis ID to a more globally persistent identifier (such as a DOI).

Continuing efforts for mature PIDs

As far as supporting mature PIDs, this is an iterative process that introduces new or additional features for further support of mature PIDs in CERN community services. Most recently, enhancements on the ORCID integration were carried out for the CERN Open Data portal (see D4.6).

Resource provenance

D4.2 focused on provenance work for FREYA pilot applications. As part of that, CERN presented developments in resource and metadata provenance. Linking publications to datasets and software has been integral for the CERN Open Data portal especially to properly highlight the components required to reuse a dataset, for example. In CERN Open Data all the published materials are carefully curated, but in CAP the users are the ones who submit information. For this case, the intention at the time was to provide a metadata provenance log from record change history information to the users through the web interface to keep track of that information more seamlessly. This task is now complete (Figure 5) and will be further improved in the future.

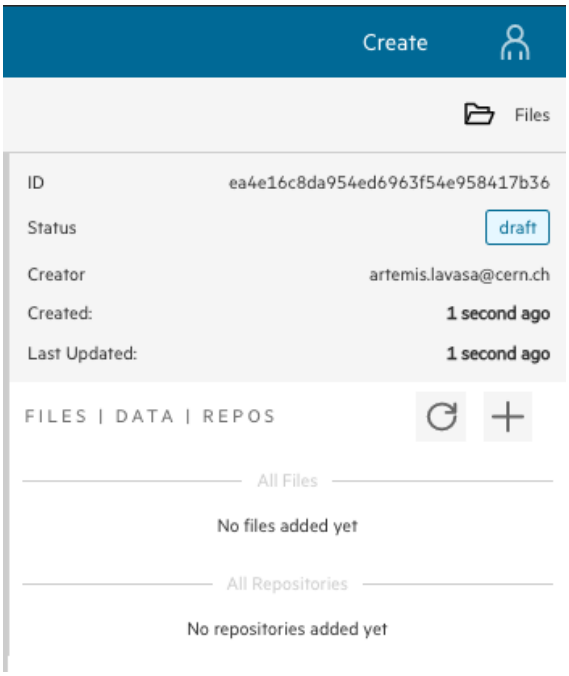


Figure 5: Tracking metadata updates in CERN Analysis Preservation records.

Outreach for PIDs within the HEP community

CERN has also done some outreach to help HEP researchers understand PIDs better and to incentivize adoption within the discipline, most notably through development of PID guides[19] with specific guidance for PIDs in CERN services, which are presented in D5.6[20]. In terms of PID guidance within services specifically, INSPIRE, for example, already had documentation for ORCID iDs primarily for users; contributors are always informed of their options with regard to including PIDs in metadata when submitting content on CERN

[19] PID guides on the CERN Scientific Information Service site: https://scientific-info.cern/submit-and-publish/persistent-identifiers

[20] Deliverable not published yet.

Open Data as everything is manually curated. For the CAP platform in particular, in order to clearly highlight for the users that they can use PIDs to enrich their analysis records and to create their own PID graph, an additional information box was added in a prominent place in the welcome page of the CAP website during a recent upgrade of the site's user interface (Figure 6).

Figure 6: Showcasing options for metadata fetching from external services using PIDs on CERN Analysis Preservation.

PID graph

The CERN PID graph (Figure 7) primarily includes DOIs for publications, data, software (including DOI versioning in some services), as well as ORCID iDs for people. Identifiers for these output types have been considered at a mature stage in CERN services since before FREYA. As mentioned in the sections above, a significant amount of work has gone towards adding more connections between PID types, and enhancing the metadata surrounding identifiers for organizations is the newest addition to the CERN PID graph and further support for ROR IDs is forthcoming.

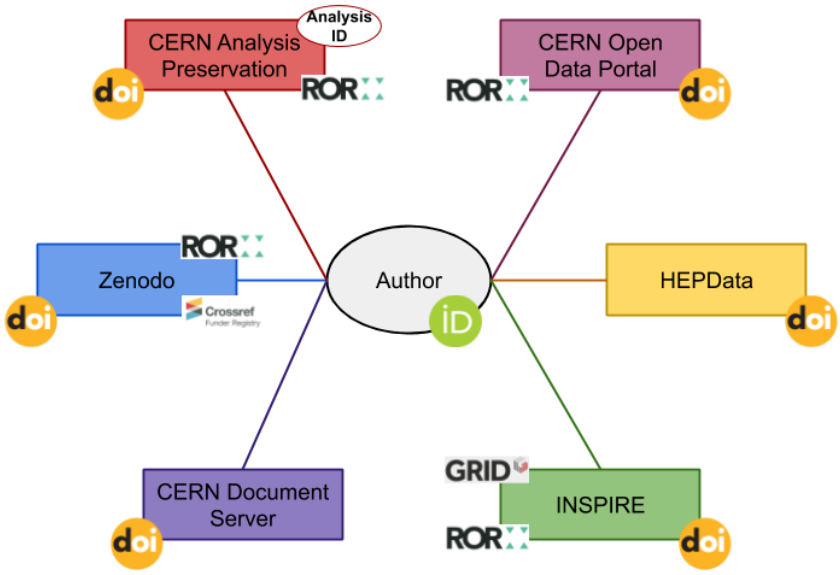


Figure 7: Diagram showing the PID types currently integrated in CERN services that include PIDs as a fundamental offering. HEPData, the CERN Document Server and Zenodo are not FREYA pilot applications; they are included in the graph for the purposes of showing a more complete picture.

Certain persistent identifiers that were noted as relevant in the first WP4 deliverable did not reach a level of maturity during the FREYA project timeline that would allow testing. For CERN services, it has always been that case that identifiers need to have reached some degree of recognition from the open science community before piloting. In terms of emerging PIDs, CERN is still interested in the PIDs identified in D4.1 as relevant for CERN, for which no work has been done as of yet, namely PIDs for instruments, conferences and workflows. Since Zenodo has already made quite a bit of progress on funder and grant IDs (see D4.6), further integration of those identifiers in other CERN services might be the logical next step.

Lessons learned

1. There is still work to be done for services in HEP in terms of integrating new and emerging PIDs, and, to a lesser degree, increasing adoption of already-integrated mature PIDs in the community. Other developments, such as the schema.org pilot that was introduced in the CERN Open Data portal (D4.1) is something that can be expanded as well in the future.
2. The CAP PID graph is an interesting use case for PIDs as it demonstrates the value of open identifiers in restricted-access systems, and the use of PIDs together with "local" and internal identifiers.
3. Too many PIDs can confuse users who might not understand the difference or possibly their value. Adding support for new PID types in the "background" can make things easier when it comes to adoption.
4. Testing out new PIDs when they are still in a pilot phase or still immature has proven to be a hard sell in that service managers are reluctant to introduce new features to users when they are not yet "stable".
5. Often, new PID integrations are put to the bottom of the priority list for development because they are not (yet) considered essential to a service and there is always another feature or fix that is needed urgently which will inevitably take priority. Such integrations are usually considered nice-to-have features and there usually needs to be pressure from somewhere for them to be added, e.g. if a funder requires it.
6. Currently, there is no one PID graph at CERN, rather a collection of local graphs that each of the aforementioned services develops. This also means that it is not yet possible to navigate from one to another through a central entry point (e.g. access a single API to get information from all CERN services using PIDs). Services are developed and maintained by different groups at CERN and an effort like this is challenging to realize, but nevertheless, this is something that would be very valuable for the community.

2.3 DANS

Pilot applications

NARCIS (National Academic Research and Collaborations Information System)[21] aggregates information from more than 48 different institutional repositories and 23 archiving systems. In addition to the aggregated metadata, NARCIS contains information from Current Research Information Systems (CRISs), information about Dutch research organizations, researchers and experts, and research projects. NARCIS also contains the project grants (project or programme descriptions) of the Dutch Research Council and European Commission (EC) as far as Dutch organizations are involved.

EASY (Electronic Archiving SYstem)[22] is an online archiving system with more than 140.000 datasets from individual researchers or research institutes. Although EASY does contain data from all disciplines, the majority of the datasets it holds are from the Humanities and Social Sciences; it contains about 100.000 datasets from the Humanities and Social Sciences. EASY is a CoreTrustSeal certified long-term preservation archive.

Technical work

Within the FREYA project DANS worked on different technical implementations to improve the PID infrastructure of NARCIS and EASY. All the following implementations are in production and available online, except for the GraphQL interface, which was built for experimentation, but not mature enough for

[21] National Academic Research and Collaborations Information System: https://www.narcis.nl
[22] Electronic Archiving SYstem: http://www.easy.dans.knaw.nl

production, and organizations and data Interlinking (see below), which is expected to be finalized at the end of 2020.

Data entry

NARCIS and EASY already supported object identifiers (DOI, Handle and URN). Within the FREYA project we added the person IDs ORCID and ISNI, and the organization IDs ROR, ISNI, GRID and VIAF. For grant IDs there is an "eu-repo namespace" and on the basis of this namespace we set up a showcase for Dutch Funders. In addition, we engaged with the NARCIS community (all universities and major research institutes) about these identifiers and the implementation in their systems.

Retrieval

- The search interface on narcis.nl provides the possibility to search for a certain PID and all objects connected to this PID will be presented, for example for 0000-0001-8879-8798[23].
- GraphQL interface (experiment, not in production).

Machine-readable data delivery

- For general purposes we realized an implementation of schema.org and JSON-LD (Linked Data), and included all available PIDs. Figure 8 shows a NARCIS page expressed in machine-readable Linked Data with PIDs.
- The PIDs in the NARCIS PID graph will be part of the metadata that NARCIS delivers to other portals and services through different metadata formats like CERIF and DataCite.

JSON-LD | POSH

Statement Collection #1	
Entity	http://orcid.org/0000-0002-8683-1854
Attributes	
rdf:type	**schema:Person**
schema:Organization	https://ror.org/01gntjh03
schema:givenName	Jaap
schema:honorificPrefix	Prof.dr.ir.
schema:identifier	***See Statement Collection #8***
schema:jobTitle	Part-time professor
schema:jobTitle	Programme leader / Group leader
schema:mainEntityOfPage	https://www.narcis.nl/person/RecordID/PRS1261211
Name	J.S. (Jaap) Sinninghe Damsté
schema:owns	https://www.persistent-identifier.nl/urn:nbn:nl:ui:10-1874-384385
schema:owns	https://www.persistent-identifier.nl/urn:nbn:nl:ui:10-1874-386382
schema:owns	https://www.persistent-identifier.nl/urn:nbn:nl:ui:10-1874-387549
schema:owns	https://www.persistent-identifier.nl/urn:nbn:nl:ui:10-1874-387640
schema:owns	https://www.persistent-identifier.nl/urn:nbn:nl:ui:10-1874-396124
schema:sameAs	http://isni.org/isni/0000000390349312, info:eu-repo/dai/nl/07401370X
Statement Collection #2	
Entity	https://ror.org/01gntjh03
Attributes	
rdf:type	**schema:Organization**
Name	NIOZ Royal Netherlands Institute for Sea Research
schema:sameAs	http://grid.ac/institutes/grid.10914.3d, http://isni.org/isni/0000000122274609
schema:url	http://www.nioz.nl/

Figure 8: NARCIS page expressed in machine-readable Linked Data with PIDs.

[23] Example of a search in NARCIS using an ORCID iD: https://www.narcis.nl/search/Language/nl/uquery/0000-0001-8879-8798

Data interlinking

All the different entities in NARCIS can be linked on the basis of PIDs. In particular, we set up special showcases for:

- Data and publication interlinking.
- Integration with ORCID where NARCIS includes relations between objects from external sources to enrich the NARCIS PID graph.
- (Funding) project and data and publications interlinking.
- Organizations and data Interlinking - not yet online, expected end of 2020.

For DOIs, there is interlinking with the external services Unpaywall, to promote Open Access, and Altmetric for information about citations on different social media platforms.

PID graph

In NARCIS as well as EASY, PIDs are playing a crucial role in identifying objects unambiguously. In addition, PIDs provide persistent accessibility and they are a prerequisite to link digital objects in a persistent way. In the communities behind NARCIS and EASY there is an increasing awareness of the importance of including PIDs in the metadata of digital objects, as they are the building blocks of any PID graph.

During the FREYA project the number of PIDs in NARCIS and EASY almost doubled and through NARCIS all these PIDs are integrated in the PID graph and accessible for other services.

Dutch community

NARCIS is organized in such a way that it enables Dutch universities and other research institutes to collaborate in knowledge infrastructure and arrangements. PIDs are an important part of these arrangements. NARCIS encourages the use of PIDs and their proper inclusion in the different metadata formats. Within FREYA and other initiatives, new PID types were promoted to the Dutch research community. In 2019 the specifications for the inclusion of PIDs were updated to help Dutch research institutes with their exchange of scientific metadata.

Person IDs

The Dutch scientific community follows a two-way policy regarding person IDs: ORCID and ISNI. An ORCID consortium tries to stimulate the use of ORCID iDs for researchers. At the same time a group of universities are implementing ISNIs for researchers, as a replacement for the Digital Author Identifier (DAI)[24], implemented back in 2008. In the national knowledge infrastructure, all three person identifiers are included in the metadata of publications and datasets and in the PID graph. Figure 9 shows a NARCIS record for a person with various person PIDs.

[24] Digital Author Identifiers: https://en.wikipedia.org/wiki/Digital_Author_Identifier

Figure 9: NARCIS person page with different person IDs

Publication and dataset IDs

DOIs and Handles are used to identify objects. Divergent is the use of URN:NBNs for all the scientific output. To guarantee that all publications (articles, book parts, theses, conference contributions, patents, etc.) contain at least one PID, all digital objects are provided with at least an URN:NBN. Future plans involve implementing a functionality to resolve the URN:NBN to a second location, namely in the Long Term Preservation depot of the National Library as a fallback mechanism for institutional repositories.

Organization IDs

NARCIS and EASY both support ROR and ISNI as organization identifiers. Although the development of an organization identifier is very promising, much outreach work will still need to be done before it will be a common practice for all Dutch Research Institutes. NARCIS provides a lot of good examples, which will encourage the use of ROR IDs and ISNIs as organization identifiers. In addition, NARCIS also supports GRID and VIAF.

Grant and project IDs

The possibility to use grant IDs has been in place for many years, but generally it is not yet used very much. Adding EC funding in the metadata of publications is mandatory for EC-funded projects and used by OpenAIRE. To increase the use of this identifier, NARCIS supports the functionality to interlink these grants with publications and data. Plans are being made by the Dutch Research Council to assign Crossref funder IDs to their grants.

Overview of the NARCIS PID graph

NARCIS ingests the above mentioned identifiers and creates links between them automatically. In numbers there are roughly: 2.150.000 objects with an URN:NBN, 850.000 with a DOI and 240.000 objects with an ISNI, including 30.000 researchers. There are 330.000 objects with an ORCID (including almost 16.000 people), 100.000 objects with a Handle and 100 organizations with a ROR ID. All these persistent identifiers are part of the PID graph and connect different information types (Figure 10).

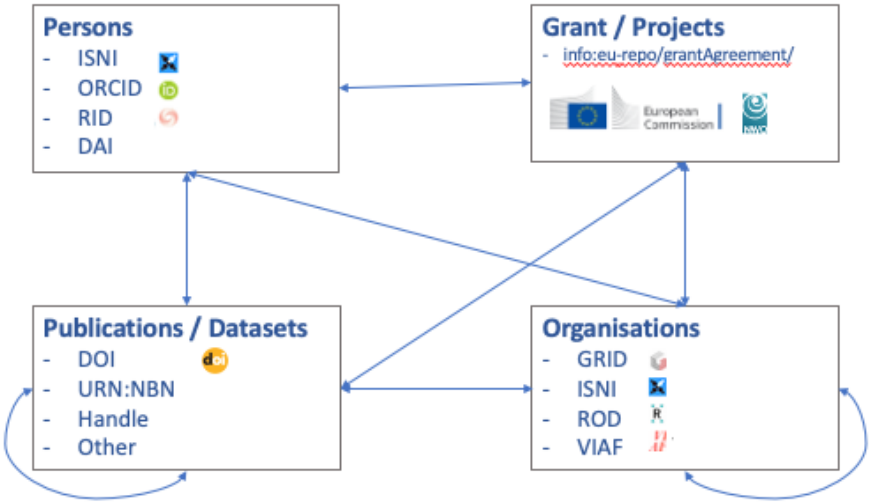


Figure 10: PID relations in the NARCIS PID graph

NARCIS uses the PID graph for the following purposes:

- Linking between objects. For instance, “all scientific outputs of a certain grant”, “all publications from a certain dataset” or “all publications and datasets of a certain researcher” can be presented by the NARCIS interface based on the PID graph.
- Exchange of PIDs with other information services. European services like OpenAIRE and B2Share, or international services like Microsoft Academic, EBSCO, ProQuest, World Wide Science use NARCIS metadata for their services. An increase in PIDs in NARCIS automatically increases the number of PIDs in these services and the relations between the different research information entities.
- Through schema.org and JSON-LD the NARCIS PID graph can also be indexed. PIDs are included in metadata standards like CERIF, DataCite, DIDL/MODS so PIDs are exchanged to other service providers.

Lessons learned

1. There is still work to be done on new PID types, e.g. for research projects. PIDs for research and funding projects can be very valuable for interlinking between objects in scholarly communication. Although there are grant IDs to identify a grant or funding project, there are no project IDs available for all research projects.
2. For projects, although some progress is made, it is to be expected that in the near future the use of grant IDs will be extended to other Dutch funding agencies and that the community will adjust their systems to support these implementations. For project IDs in general there is no solution yet. Probably smaller subsidiaries will follow these examples, but more outreach to these institutes is needed.
3. The real challenge of a new PID type is twofold: creating a PID agency and stimulating the use of the PID in different communities with different metadata standards, applications and workflows, which takes time and outreach work.
4. Maybe there are too many PIDs for one type from a user perspective. The discussion about DOI vs Handle, ORCID vs ISNI or ROR vs ISNI will remain. The PID Graph could be a strong tool to (partly) solve this problem when it can identify the same object using multiple identifiers.
5. The use cases for a PID Graph in the Social Science and Humanities are mostly related to very concrete questions by researchers and funding agencies: e.g. “all data and publications from a certain grant” or “which data is connected to which publication”.

2.4 EMBL-EBI

Pilot applications

Europe PMC, one of >40 repositories hosted by EMBL-EBI, is a literature database that aggregates biomedical journal articles and preprints. Article records are enriched with links to underlying data sources and resources such as ORCID iDs, open citations, text-mined concepts and data citations, and grant information. These links are PIDs that direct users to the specific resources. The content in Europe PMC is updated daily and can be searched and retrieved via the website[25] and APIs[26].

OmicsDI[27] aggregates genomics, transcriptomics, proteomics, metabolomics and multiomics datasets, as well as computational models of biological processes. On publication, data from multi-omics studies often have to be deposited in multiple technology-specific repositories. Based on the PMID, OmicsDI aggregates metadata from many omics data resources, providing an integrated view of all data referencing a specific publication. To contribute to the valuation of datasets in their own right, OmicsDI also aggregates multi-dimensional usage data for its entries. Specifically, metadata views within OmicsDI, data download statistics from the source data repository, citations of the actual dataset accession number in Europe PMC, and references to the dataset in curated knowledge bases like UniProt are gathered, aggregated, scaled, and provided in a user-friendly form.

Technical work

As part of EMBL-EBI's contribution to Work Package 4, various existing Europe PMC services that use PIDs were enhanced with more features and functionalities.

Integration of mature PID types
Mature PID resources in Europe PMC include PIDs for publications (DOIs, PMID, PMCIDs), authors (ORCID iDs) and data (data accession numbers, DOIs).

Data accessions and DOIs
Additional data accessions were introduced which is of relevance to research relating to the coronavirus pandemic of 2020. Text mining has been used to expose the GISAID data accession numbers in publication records, the metadata of which provide genetic sequence, and related clinical and epidemiological data associated with human viruses. Where present in the publication records, these GISAID data accessions are highlighted (see Figure 11) and thereby linked into Europe PMC's PID graph.

[25] Europe PMC website: https://europepmc.org
[26] Europe PMC's developer resources: https://europepmc.org/developers
[27] OmicsDI: https://omicsdi.org

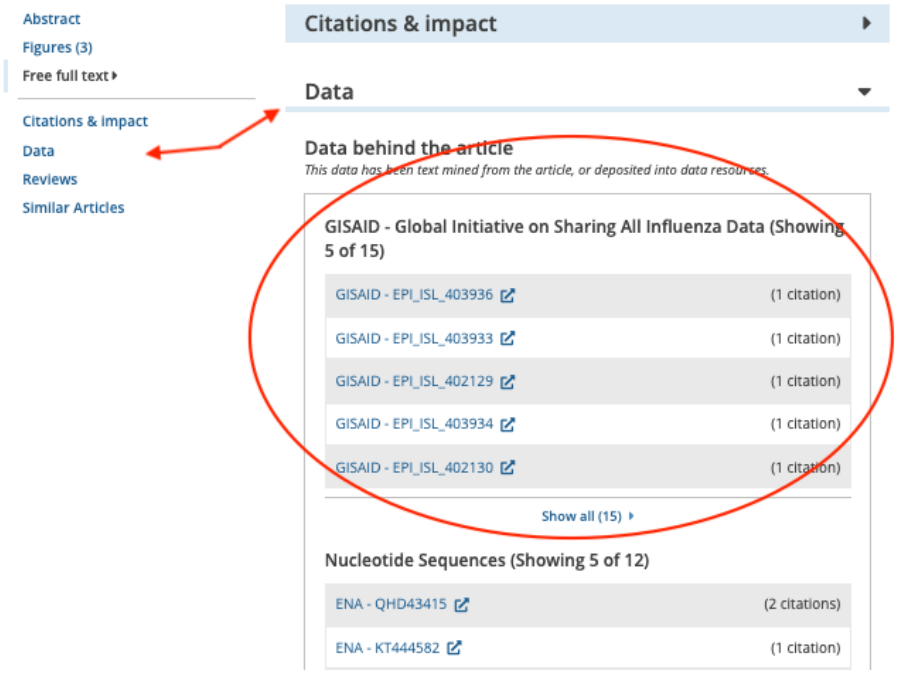


Figure 11: A screenshot to show the GISAID links in the data section of an article indexed in Europe PMC. A reader is directed to the original resource for this data record. This aids in data discovery and provenance. Articles containing GISAID accessions can be found in Europe PMC using the boolean query "ACCESSION_TYPE:gisaid"

Another piece of work relating to DOIs involved data cleanup. A small proportion of the data PIDs in Europe PMC are data DOIs. Given that DOIs are also used for publications, DOIs can be incorrectly recorded in the literature and therefore incorrectly linked into Europe PMC's PID graph. A clean-up project is underway to check the set of data DOIs in Europe PMC and ensure that these are correctly "tagged" to datasets. The work involves development of an algorithm that recognises data DOIs in Europe PMC's content, checks the integrity of the link to data using information in DataCite's DOI checker and then retains the link if the "resource-type-subtype" is associated with a "dataset", but deletes the tag if it is associated with "article" or "software" for example. The algorithm will be deployed in Q4 2020.

Provenance of content

To enrich the provenance information available for the growing biomedical content indexed by Europe PMC, the team focused on exposing versions of preprints for users and linking these to journal published versions when they appear. The preprint versioning implementation in Europe PMC has been extended since D4.3. The mechanism to identify and handle versioning in Europe PMC was extended to versions of full-text preprints. While the full-text preprint indexing initiative itself was funded through sources other than FREYA, making the codebase for handling versioning of these preprints open source was FREYA work.

Identifying and handling preprint versions via Europe PMC's manuscript submission system

In 2020 Europe PMC has developed a new route for ingesting the full-text of COVID-19 preprints - via Europe PMC plus, Europe PMC's manuscript submission system (MSS) through which journal-accepted articles are ingested. In addition to MSS changes to handle preprints as a new content type, Europe PMC has made additional changes to accommodate versions of preprints, and to correctly match PIDs in order to create the best experience possible for our users.

All the preprint servers that post the life sciences preprints indexed in Europe PMC allow the creation of multiple versions of a preprint, to track changes made. These versions are assigned identifiers in different ways by different preprint servers, leading to different combinations of PIDs for each preprint and version. We have identified two main ways preprint servers register preprints with DOI registration agencies, namely, versions with a shared DOI and versions with different DOIs.

The first time any version of a preprint is received by the MSS and ingested, its PIDs are checked against already recorded PIDs for any duplication errors, after which it is processed by the system for the creation and approval of XML plus associated web versions for loading into Europe PMC's database. The loading of additional versions received by the MSS depends on the PIDs assigned to the preprint and its versions.

The newly altered code base for Europe PMC's manuscript submission system has been moved to the open GitLab project (maintained by the Coko Foundation[28]) to make it freely available for reuse[29].

Organization IDs

Although the infrastructure is not as mature as for publications, authors and data, ROR IDs and the ROR repository emerged at the outset of 2019 enabling FREYA partners to conduct early pilots for WP4. To identify publications by EMBL-EBI authors, Europe PMC has worked specifically on an initial mapping project (D4.4) and then more refined machine learning approaches that allow for ROR IDs to be assigned retrospectively to EMBL-EBI authors listed on existing Europe PMC records (see D4.6). Conference presentations have been undertaken (at Biocuration 2019, HUBS 2019, JATS CON2019 and RRTS20) that explain to researchers the benefits of including organization IDs in their manuscripts from the outset.

Identifying research resources in the biomedical literature without mature global PID systems

Building links between publications and research resources such as funding information and publication licenses provide users with information that, for example, helps them to fulfil requirements for the Research Excellence Framework (REF) exercise. The REF is the system for assessing the quality of research in UK higher education institutions that runs every 5 years in the UK. To this end we are exploring methods to identify missed funding or licensing resources and ensure these are "tagged" in relevant publications. Specifically, this involves determining how to extend the current integration with Crossref to extract additional funding information and licensing information. Efforts like this ensure the crosslinks in Europe PMC are as complete as possible and serve as a starting point to contribute to a global infrastructure for licensing and to build these funding acknowledgements into the emerging Global Grant ID system.

PID graph

As mentioned above, the mature PID resources that are included in Europe PMC are publications (DOIs, PMID, PMCIDs), authors (ORCID iDs) and data (Data accession numbers, DOIs). Europe PMC's biomedical publication content increases daily. Biomedical preprints from an increasing number of servers now form part of the core content - including links to the full-text provided by DOIs. Article records in Europe PMC are enriched with links to underlying data and other research resources. Figure 12 presents the scale of the interconnected PIDs in Europe PMC's "local" graph, representing a rich resource for researchers.

[28] Coko Foundation: https://coko.foundation/

[29] Public manuscript submission system project (Europe PMC plus): https://gitlab.ebi.ac.uk/literature-services/public-projects/xpub-epmc

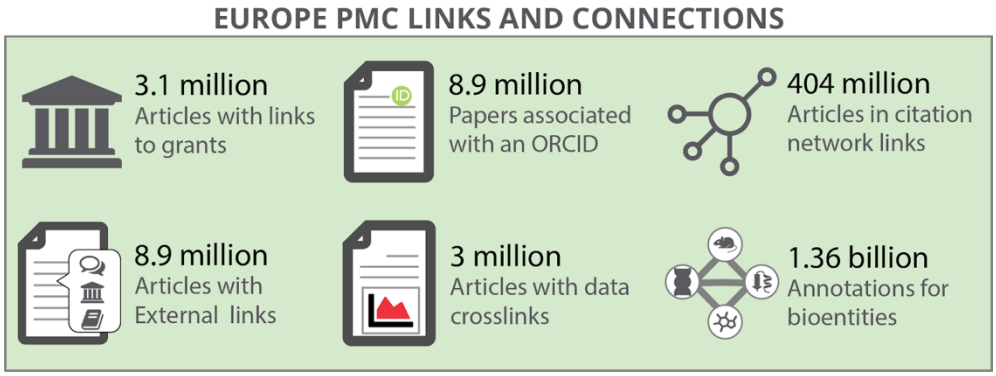


Figure 12: Article records in Europe PMC are enriched with links to underlying data and other research resources. This is the status of Europe PMC's links and connections in Q3 2020. Mature PIDs are used in links involving people (ORCID iDs), data crosslinks and article citations. Grants, external links and annotations are not fully supported by mature PIDs (see the main text below for further explanation).

The proportion of articles that are linked to ORCID records increases year on year (57% of abstracts from PubMed are now linked to at least one ORCID record) and over 66.000 preprints have been linked to researchers' ORCID records. This is in part due to incentives, such as introducing functionality that makes it easier for an author to claim articles to their ORCID profile. Over 3 million publications now link to related data: Europe PMC has worked to increase the types and numbers of data crosslinks available and exposed this for users in a redesigned data section for each article.

As far as grants are concerned, Europe PMC has a grants database[30] that holds funding details for grants awarded by its 31 European research funders[31] so that articles and grants can be linked. Global grant IDs (DOIs) are not discussed here since DOIs for grants and the associated infrastructures are currently emerging. To date, only several hundred DOIs for grants have been registered by Crossref. The grant identifiers (mentioned in Figure 12) are not global identifiers, but internal identifiers devised and assigned to grants independently by each funder.

Currently, 8.9 million articles in Europe PMC feature external links; these represent collaborations between Europe PMC and nearly 60 different providers that link research articles to useful resources providing free access to peer reviews, recommendations, protocols and materials, lay summaries, etc. Only some of these are assigned DOIs; others have identifiers, but not necessarily global identifiers. Annotations are biological entities and concepts, such as data, genes/proteins, diseases and experimental methods that are text mined from articles in Europe PMC. Europe PMC's annotations platform collates the text-mining results produced by the Europe PMC team as well as those submitted by the external text mining community. Not all of these have mature PIDs.

Additional user stories that can be addressed using Europe PMC's content and services, and contribute towards Europe PMC's PID graph, include: retrieving publications that refer to datasets from data repositories and thereby use the citations of those papers to get a sense of how much a dataset might be reused; this is addressed by Europe PMC's text mining of data accession numbers. Although not developed FREYA funding per se, Europe PMC has a community platform whereby community text miners can contribute their text mining efforts using Europe PMC content and retrieve it too. A network analysis of the researcher's collaborations can be achieved using Europe PMC's Articles API and ORCID integrations to retrieve the co-authors of a particular researcher. For example, OmicsDI uses Europe PMC's API to mine citations of dataset accession numbers, providing one element of an aggregated prototype dataset impact metric.

[30] Europe PMC's grants database: https://europepmc.org/grantfinder
[31] 31 research funders of Europe PMC: http://europepmc.org/Funders/

Lessons learned

1. Europe PMC is a repository hosting content from many publishers and preprint servers, and provides a search engine for this content working off PID graph connections. Archiving and linking this content is made possible because of PIDs and metadata standards. Nevertheless, a repository is confronted with connection challenges where sources deviate, e.g. providing solutions to accommodate the variety of methods employed by sources to mark versions (mentioned above), or the manner in which retractions/withdrawals of research publications are handled. These impact on the accuracy of the connections made between research resources and the provenance.
2. Challenges of including new research resources: DOIs are registered for journal published articles as well as for preprints. Yet PID standards agreed by publishers are more advanced than those for preprint servers. This requires the engagement of source communities, feedback at community meetings and flagging issues. To this end, Europe PMC has presented FREYA outputs at publisher meetings (OASPA 2019, 2020, CISPC 2019), data meetings (FORCE 2019), and at PIDapalooza 2019, 2020. EMBL-EBI has also hosted a workshop for preprint server staff to agree on a number of standards (not FREYA funded)[32].
3. Where different PIDs are used for a single resource: such as for data in the biomedical sciences where data accessions are very commonly used (compact identifiers that comprise any local unique identifier with a prefix that is "repository identifying"[33]) and DOIs to a lesser extent. Use of identifiers by a discipline will vary for a specific research resource. These bring different challenges such as cleaning up the metadata (mentioned for DOIs above) or mapping resources for which more than one identifier is used.
4. The content within Europe PMC's PID graph is far from saturated and can be grown by constantly encouraging content generators to incorporate PIDs in their work (in this case life science researchers who generate data and publish their research). Incentives for researchers can include publishing workflows that generate publications by incorporating data PIDs and metadata directly[34], and advocacy by leading researchers, for example at conferences[35].
5. To broaden the reach of crosslinking and to address additional user stories or complex queries, Europe PMC and other repositories can draw on the Jupyter notebooks and GraphQL resources reported below.

2.5 PANGAEA

Pilot applications

PANGAEA's work in WP4 has focused on populating dataset metadata of PANGAEA records with PIDs so that all research components are linked in a reciprocal manner. Efforts have gone into enriching metadata to include grant IDs, organization identifiers (ROR IDs), and Handles for instruments, addressing use cases for funders and organizations with regard to output and impact of funded research.

We also used the development of a sample IGSN App (D4.3) to demonstrate how the implementation of the different PIDs in the dataset metadata can be navigated by users to reconstruct the research environment and to discover additional data (DOIs), research (DOIs, ORCID iDs) and samples (IGSNs) related

[32] ASAPbio January 2020 workshop: A Roadmap for Transparent and FAIR Preprints in Biology and Medicine: https://asapbio.org/meetings/preprints-roadmap-2020

[33] Wimalaratne SM, Juty N, Kunze J, Janée G, McMurry JA, Beard N, Jimenez R, Grethe JS, Hermjakob H, MartoneME, Clark T. (2018): Uniform resolution of compact identifiers for biomedical data. Sci Data [08 May 2018, 5:180029]: https://europepmc.org/abstract/MED/29737976

[34] For example Datascriptor: https://datascriptor.org/

[35] Advocacy at OASPA 2020: https://twitter.com/ChrisVFerg/status/1308821054429814784; advocacy at RRTS20: https://twitter.com/ChrisVFerg/status/1304074985485590529

to their own work. IGSNs (International Geo Sample Number) are PIDs that uniquely identify samples from the natural environment and related sampling features; they are minted by different designated allocating agents which follow common standards and procedures. Additional work has gone into improving the connectivity within the IGSN hierarchical architecture to ensure navigation is possible from all levels of the hierarchy, starting from the ship expedition→ sampling sites → bore holes → cores → samples → subsamples, or any other part of the tree and in any direction.

Technical work

Sample PIDs (IGSNs)

Earlier work on integrating IGSNs (sample PIDs, D4.3) as part of dataset metadata has now been expanded to establish sufficient connectivity between the different hierarchical levels within the IGSN architecture. The PANGAEA/University of Bremen's FREYA team worked together on a development sprint with the IGSN2040 project[36] to enrich the landing pages of IGSN persistent identifiers to contain machine-readable JSON-LD metadata in different formats (schema.org, but also custom metadata profiles to describe metadata of samples and their relationship between drill holes, sites and ship expeditions/legs. For the IGSNs minted by the University of Bremen/MARUM, PANGAEA added a mockup implementation of those landing pages (see Figure 13). In the near future they will replace the database-generated pages by the DIS software[37] used by the MARUM/IODP core repository.

The problem of the current DIS system is the inability to implement landing pages easily, as the system does not allow adding separate pages for every IGSN. To implement JSON-LD/schema.org metadata, every IGSN must have their own landing page.

BCR Core Repository Bremen - DIS View: Core-Details

Go to repository login-page: Repository login

Core:

Expedition	Site	Hole	Core	Type	Top Depth (m)	Drilled Length	Bottom Depth	MCD Offset	Recovery	Recovery(%)	Oriented	Sections	Core Catcher	Curator	IGSN	Comments
302	1	A	1	H	0	0.31	0.31	0	0.31	100	no	1	yes	AW	IBCR0302RCV6G01	IWRH @

Sections:

Section	Section Length	Curated Length	Top Depth (m)	Bottom Depth (m)	Curator	CC	MCD Top	Box	Slot	Position	IGSN	Comments
1	0.32	0.37	0	0.37	AW	yes	0	0			IBCR0302RSG1BS2	

Measurements, descriptions and images:

Section	Section Desc.	VCD-File	Slabbed Scan
1			

Samples:

Section	Half	Type	Top (cm)	Bottom (cm)	Top MCD (m)	Volume (cc)	Request	IGSN	Sample	Comments
1	W	PAL	2	4	1.02	10	MSP9999	IBCR0302RXBONS2	3016664	Section 1 = CC
1	W	PAL	2	3	1.02	0	MSP9999	IBCR0302RXMPNS2	3016714	Section 1 = CC
1	W	PAL	10	12	1.1	10	MSP9999	IBCR0302RXBONS2	3016667	Section 1 = CC
1	W	---	11	13	0.11	10	037251IODP	IBCR0302RX48KZ3	5018404	
1	W	PAL	14	16	1.14	10	MSP9999	IBCR0302RX9ONS2	3016665	Section 1 = CC
1	W	PAL	16	18	1.16	0	MSP9999	IBCR0302RXNPNS2	3016715	Section 1 = CC
1	W	PAL	18	18	1.18	0	MSP9999	IBCR0302RXKPNS2	3016712	Section 1 = CC

Figure 13: Current landing page of the MARUM DIS system. Currently, multiple samples are embedded in one HTML page, making it impossible to embed sample-specific JSON-LD metadata.

In the future, all IGSNs will resolve to the new landing pages and only display and format the metadata from the proprietary DIS back-end database. Currently this information is not easily available, as DIS software, which is provided by a third party, is intended only for internal use and allows no customizations.

[36] IGSN 2040 (funded by the Alfred P. Sloan Foundation): https://www.igsn.org/igsn-2040/

[37] Conze R. (2016): Drilling Information System (DIS) and Core Scanner: http://doi.org/10.17815/jlsrf-2-130

The FREYA IGSN app cannot “walk” through relationships between samples, holes, sites and legs, because the information is not machine-readable.

The new mockup implementation created as part of the IGSN2040 sprint provides a separate landing page for every IGSN identifier and includes machine-readable metadata in JSON-LD format (Figures 14 and 15). This allows e.g., PID graph services to navigate through the graph starting from an IGSN sample identifier up to drill holes and drill sites.

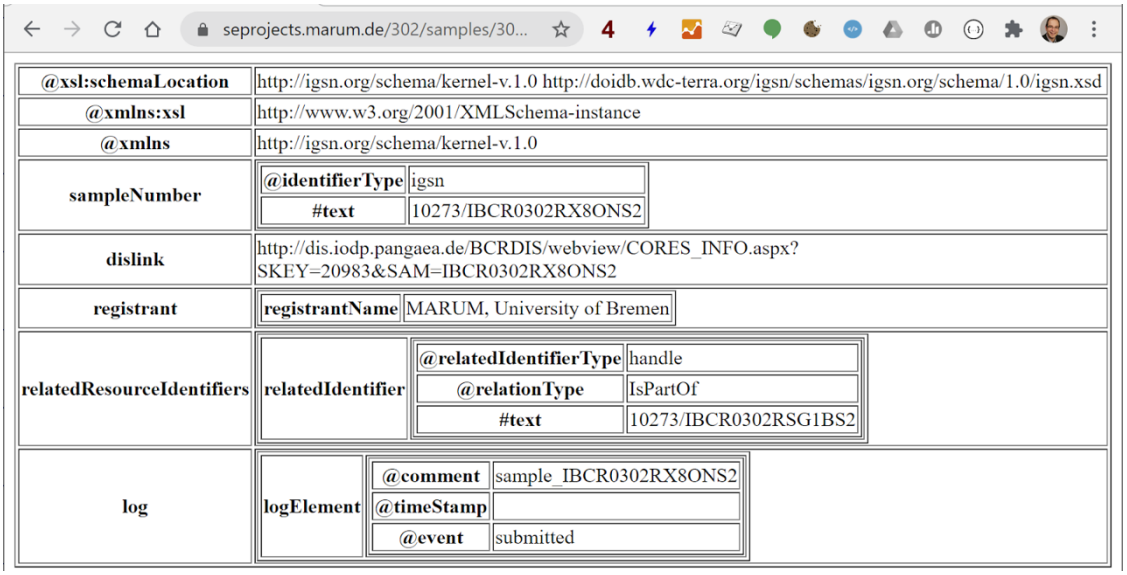


Figure 14: Mockup of a new IGSN sample landing page for a series of related samples. The landing page is at an early stage visually, merely serving as a mockup to demonstrate what can be done.

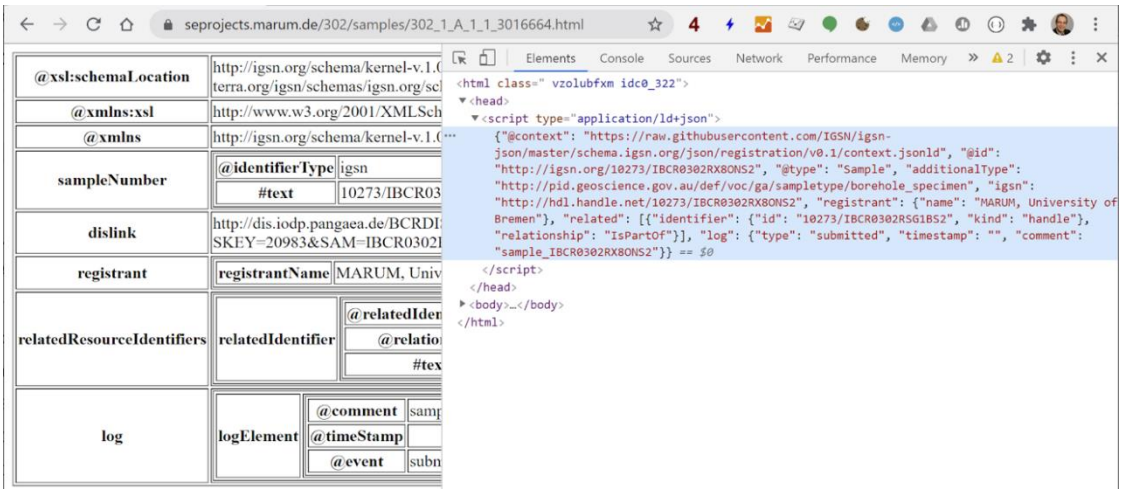


Figure 15: JSON-LD metadata embedded into the mockup landing page

A GraphQL application can resolve the IGSN identifier (e.g. for a sample), read the metadata from the landing page and walk up the hierarchy to get metadata of drill holes or sites. This allows the FREYA IGSN barcode app to create the link between samples and datasets: e.g. PANGAEA only links drill holes in their metadata, but not individual samples. If someone scans a barcode of a sample, the link to datasets was therefore missing. Now it is possible to follow the PID graph to link samples with datasets.

The implementation was made available as a mockup and will be implemented in the future once the full metadata profiles are negotiated in the IGSN2040 project. Source code of PANGAEA’s mockup

implementation for the MARUM DIS system is provided on GitHub[38]. To allow harvesting of all landing pages of MARUM-registered IGSNs a sitemap was provided[39]. The mockup implementation (a Python script to generate HTML pages and sitemap from the IGSN XML metadata used for registration, exported from the DIS database) was deployed on a MARUM test server. At a later stage, the landing pages will be formatted to be more human-friendly and registered as the main landing page of all IGSNs registered by the MARUM implementation.

The FREYA IGSN App will be adapted to use the machine-readable metadata to navigate the PID graph and will mainly use a JSON snippet to do so (marked red in Figure 16 below):


```
{
 "@context": "https://raw.githubusercontent.com/[...]/context.jsonld",
 "@type": "Sample",
 "@id": "http://igsn.org/10273/IBCR0302RX8ONS2",
 "additionalType":
"http://pid.geoscience.gov.au/def/voc/ga/sampletype/borehole_specimen",
 "igsn": "http://hdl.handle.net/10273/IBCR0302RX8ONS2",
 "registrant":{
  "name": "MARUM, University of Bremen"
 },
 "related":[
  {
   "identifier":{
    "id": "10273/IBCR0302RSG1BS2",
    "kind": "handle"
   },
   "relationship": "IsPartOf"
  }
 ],
 "log": {
  "type": "submitted",
  "timestamp": "",
  "comment": "sample_IBCR0302RX8ONS2"
 }
}
```


Figure 16: Excerpt of the machine-readable metadata schema for IGSNs registered through MARUM/University of Bremen. The FREYA IGSN App pulls resources from the schema using the JSON snipped highlighted in red.

At a later stage a GraphQL implementation will be provided as well. For this, the sitemap and JSON-LD embedded on landing pages can be harvested to build a PID graph in a database, e.g. located at IGSN organization.

Organization identifiers (ROR)

The inclusion of ROR IDs for organizations is ongoing work and is reported in more detail in D4.4 and D4.6. We used the ROR API to provide a first matching between PANGAEA's internal organization registry and the ROR registry, which then had to be further manually scrutinized for false negative and false positive results. ROR IDs in PANGAEA datasets have a user-facing implementation in cases where research organizations are part of the dataset title and also as part of the metadata describing projects (included for the coordinating organization of the project). While PIDs for projects (e.g. RAiD) are not yet in wider use, project metadata collected now contain quite a few interlinked PIDs: Coordinator (ORCID), Affiliation (ROR), Funder (Crossref funder ID), and a Grant ID if funding comes from the European Commission or the Deutsch Forschunsgemeinschaft (DFG), in addition to linking to project web pages.

[38] MARUM DIS IGSN landing page mockup implementation on GitHub: https://github.com/pangaea-data-publisher/marum-dis-igsn

[39] MARUM DIS Sitemap: https://seprojects.marum.de/sitemap.xml

Grant IDs

We integrated grant IDs (not DOIs) to enrich project-related metadata, by including actionable links to landing pages provided by funders (European Commission funding lines and DFG). The grant IDs will be replaced by PIDs (grant DOIs) as soon as they become available.

Data provenance (including instrument Handles)

Provenance information about PANGAEA datasets (e.g. changes to dataset metadata, versioning) can be retrieved via the DataCite Provenance API[40], in addition to the already built-in option of "is version of" or "is newer version of" linkage in PANGAEA metadata. In the context of the PANGAEA infrastructure, we also consider information about how the data was collected as provenance and therefore include instruments for data and sample collection. Handles minted for instruments by sensor.awi[41] are already included in the metadata, if they are provided by the user. The MOSAiC expedition, which has just returned to harbor this month (October 2020), is considered a turning point in the use of instrument PIDs, and a switch from institutional Handles to minting of DOIs and ePIC Handles[42] for instruments is anticipated shortly, as data from this year-long expedition into the Arctic will start making its way to PANGAEA for publishing.

PID graph

The PANGAEA PID graph has grown from including only PIDs for authors, articles and datasets to also including newer PIDs for instruments, organizations, physical samples and grants (Figure 17). Future work will continue to focus on linking these PIDs to improve the users' ability to "enter" the graph from any node, leading to improved discovery and reuse of data within PANGAEA's infrastructure. PIDs and machine-readable metadata are a prerequisite for efforts to evaluate the "FAIRness" of data in data archives. New tools to accomplish this task with machine-to-machine communications are being developed and can soon help users determine the quality of metadata provided with the data offered, helping to drive the reuse of data. Improvements of the graph lie in building a tighter network of connections, so that users can achieve more complex queries using the graph and external resources.

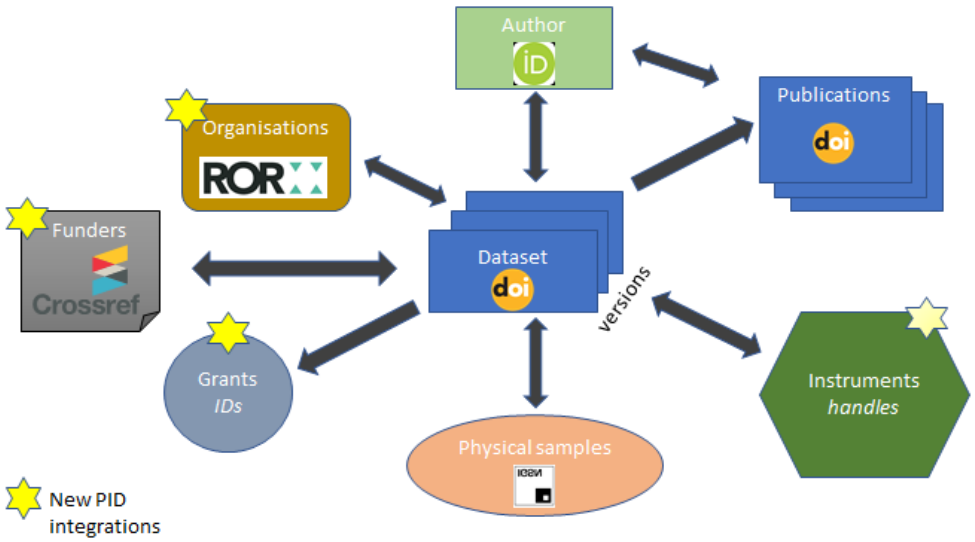


Figure 17: The PANGAEA PID graph within the data architecture. Additional ways to navigate the graph and its nodes include building applications like the IGSN App developed to provide users of the Bremen Core repository improved access and discovery.

[40] DataCite's API for tracking metadata provenance: https://support.datacite.org/docs/tracking-provenance
[41] Sensor.awi: https://sensor.awi.de/
[42] ePIC: https://www.pidconsortium.net/

Lessons learned

1. Incentivize adoption: PANGAEA serves a large research community for their data archiving and publication needs in a cooperative manner. Including persistent identifiers with machine-readable metadata as part of dataset publications provides greater ease during data publication (for users and curators) and even greater benefits for users in search of data or other information. However, the incentive to provide the information can often be relatively low. To acquire an external PID (e.g. for your instrument) takes additional time and effort and is not mandatory for publication. Providing users with tools (e.g. the presented Jupyter notebooks and GraphQL resources) to benefit directly from the inclusion of PIDs in metadata can motivate researchers to invest more time. Also, providing up-to-date resources pertaining to PID use (PID Forum) helps reduce barriers to adoption as well.
2. Among the greatest challenges for a more broadscale use of the IGSN PID and expansion to other disciplines, is to ensure the scalability of services and technical infrastructure. The IGSN registration is provided by a federation of IGSN Allocating Agents, which manage this task according to the very specific needs of their community (often single institutes). More work is needed to ensure better integration of formats and machine-readability of metadata content (includes development of APIs). This will also improve the discoverability of samples through associated research aspects (e.g. sample site, cruise leg, bore hole, data publication).
3. PANGAEA supports community adoption of new PID types simply by including these as options in dataset metadata. Due to the one-on-one support during data submission and publications, curators can often convince authors to make the extra effort to research the needed PIDs (for example, to provide ORCID iDs for all co-authors). Through PANGAEA's participation in numerous past and ongoing research projects, we educate and encourage consortia to make common commitments to the use of a core set of PIDs to ensure that data publications meet minimum standards.

2.6 STFC

Pilot applications

STFC is developing the Open Science Portal - a new service that integrates records of science from several institutional repositories and links them to quality information resources beyond the organization walls. The role of persistent identifiers in this emerging service is twofold: first, already-assigned PIDs are a means of integration for the records of science dispersed across different information sources and second, more PIDs can be introduced (assigned) in certain cases as a means of enrichment of the existing records of science (augmenting what is now the uncontrolled free-text, e.g. organization names, with clear PID associations).

The baseline for these works was the knowledge graph for PhD theses and related entities reported in D4.3. That initial graph contained only a few hundred nodes and relationships but was further expanded and helped to develop a particular user story about PhD research outcomes, as well as a sensible approach to the integration of metadata coming from a variety of sources, also to test the suitability of certain

technological solutions. These foundational works have been performed in collaboration with the British Library and resulted in a number of conference presentations and publications[43,44,45].

The graph has now grown both in size and in complexity and contains about 90 thousand nodes representing various STFC records of science including but not limited to the PhD theses, and about 12 thousand relationships across these records of science and further to external records, exemplified by records in the Cambridge Structural Database and Protein Databank. The works on the back-end database have been accompanied by a more prominent effort for the development of the user interface and a wider engagement with STFC stakeholders, to secure the sustained effort on the pilot using STFC's own resources beyond the lifespan of FREYA.

Technical work

The focus of the STFC Open Science Portal work has been on building a graph database and a web interface to it, as well as a number of demo web pages that illustrate particular user stories and serve as a "laboratory" to collect users' feedback and decide on the future directions of the Portal's development.

Technologically, STFC Open Science Portal is being implemented as a multi-tier web application with the graph database back-end and ETL components for data acquisition. The web tier is built with Python libraries and JavaScript components. The current scale of the database is about one hundred thousand records (nodes) and over ten thousand relations (edges). There is a clear potential to scale up both the number of nodes and number of relationships, specifically with the inclusion of citations and links to Open Access versions of research papers.

The Portal allows indexing and searching for the STFC records of science irrespective of their nature: publications, datasets, funding, and showing the cross-links among these records, as well as links to the external information sources of a reference quality such as the British Library EThOS service for theses, Cambridge Structural Database for crystallographic data records and Protein Data Bank. The integration with the Protein Data Bank is small-scale at the moment (a few hundred records) but there is an ongoing joint effort with EMBL-EBI to scale it up to thousands of records.

To advance the front-end development, STFC stakeholders have incrementally developed a range of user stories that were then passed to a subcontractor for implementation using JavaScript libraries. These implementations will be used in a "Show-and-Tell" event planned for STFC stakeholders and FREYA partners; the feedback collected will help to set priorities and select technology for the front-end development of the Portal.

The visualizations, along with less user-oriented techniques, can be used for research impact studies tailored to STFC facilities needs and in many cases, they rely on sensible record aggregations. PIDs are a key to the automated aggregations used in the analysis of facility research outcomes. Figure 18 below illustrates an evolution of the Open Access rates for the research outcomes of a particular facility instrument. The data used in this illustration has been collected as a joint dedicated effort of the instrument scientists and bibliographers. With the wide adoption of instrument PIDs, data aggregations of this kind can be automated, and preparatory work has been carried out for this by disambiguating research instrument names as reported in D4.6.

[43] Bunakov V., Madden F. (2020): Integration of a National E-Theses Online Service with Institutional Repositories. Publications 2020, 8(2), 20: https://doi.org/10.3390/publications8020020

[44] Madden F., Bunakov, V. (2019): Using persistent identifiers to track PhD outcomes. Presented in ETD 2019: Fruits of Knowledge: https://www.bad.pt/publicacoes/index.php/cadernos/article/view/2020/0

[45] Bunakov, V. (2019): Metadata Integration with Labeled-Property Graphs. In Metadata and Semantic Research. Communications in Computer and Information Science 1057 edited by E Garoufallou, F Fallucchi, E William De Luca, 441-448. Cham: Springer International Publishing: http://dx.doi.org/10.1007/978-3-030-36599-8_41 Open Access version: http://purl.org/net/epubs/work/44669470

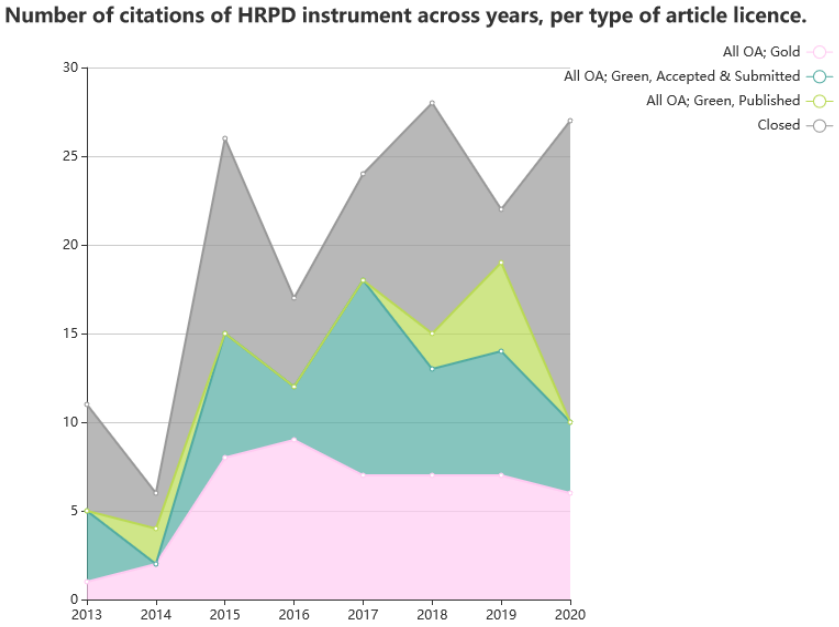


Figure 18: Rates of Open Access research outcomes for the HRPD instrument on the ISIS facility[46]

There is an emphasis on interactivity and animation for the visuals, which should allow richer storytelling with data. The visuals for the user stories will be initially incorporated in the Portal as separate web pages (one-page web applications), then after collecting the stakeholders' feedback, the most compelling visuals will be integrated with the Portal's search functionality.

The STFC Open Science Portal may not be only a consumer but a provider of PID graphs, too, through the exposure of records and their interconnections via the GraphQL API. Then the Portal becomes a part of the virtuous circle when it can consume certain parts of a PID graph and contribute to it, too. Technological experiments with the GraphQL API provision have started and will be continued beyond the FREYA lifespan.

Finally, a discussion with ORCID has indicated a potential for the disambiguation of authors for a certain number of research papers and datasets already acquired by the Portal. Doing this at scale requires a dedicated effort and will be opportunistically pursued beyond the lifespan of FREYA.

PID graph

The STFC PID graph involves:

- DOIs for research papers in STFC repositories and for the Open Access versions matched through the Unpaywall service.
- DataCite DOIs that contain metadata of facility experiments (investigations) with further links to the raw data collected in the experiments.
- DOIs for "long-tail" datasets in the STFC data repository.
- DOIs for the Cambridge Structural Database records with crystallographic data matched to the research papers.
- ROR IDs for organizations (mostly UK universities) funded by STFC for their projects or studentships.

Apart from the above listed entities identified by PIDs, the back-end graph contains disambiguated records for:

- STFC funding (grants).
- Projects and studentships funded by STFC.

[46] High Resolution Powder Diffractometer: https://www.isis.stfc.ac.uk/Pages/hrpd.aspx

- Large-scale instruments (facility beamlines).
- Protein Data Bank records - these disambiguated records have a potential to be further assigned with existing or new types of PIDs when research community practices and technology are ready for it.

The disambiguated entities should be seen as the initial step towards further adoption of persistent identifiers for them. This stance, as well as challenges for the adoption of new PID types are explained in more detail in D4.6.

The following figure (Figure 19) gives an example of a subgraph extracted from the Open Science Portal, with some nodes having associated PIDs, and other nodes just disambiguated (awaiting PIDs):

Figure 19: Publication in the Diamond bibliographic database (central node) connected to its Open Access counterpart in Europe PMC (top right), to the Cambridge Structural Database record (bottom right) and to three Protein Data Bank records on the left. The metadata sources from where these records are harvested do not necessarily "know" about each other, but their integration in the STFC Open Science Portal provides connections and allows crosswalks between any of the records of science involved.

The focus for the Open Science Portal for the foreseeable future is going to remain on the metadata and its visualization, and data visualizations are going to be incorporated where effort to produce them is moderate. Figure 20 contains a visualization of a dataset from STFC's "long-tail" data repository.

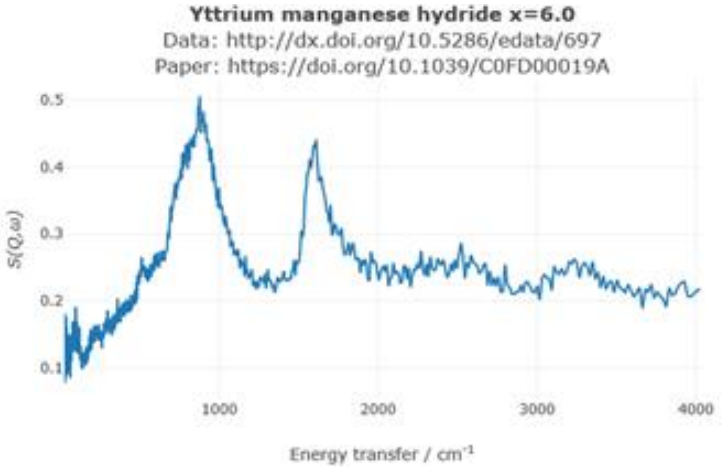


Figure 20: Visualization of a dataset from INSDB (Inelastic Neutron Scattering Database). The visualization is interactive when accessed in a web browser and includes DOI-based URLs on the top that refer to the data record in INSDB and to the paper that used the data, so that the visualization gives a good idea of both what the data is and where further detail of it can be found through the resolution of persistent identifiers.

The work on the Portal prototype has been communicated in two online events[47,48]. The software code of the prototype and the examples of visualizations are going to be published on Zenodo after the online Show-and-Tell event planned to take place in November 2020.

Lessons learned

1. The natural difficulty of using different sources for feeding into the STFC Open Science Portal is the diversity of their APIs, so these works are being prioritized depending on the value that the integration of a particular information source brings to the Open Science Portal pilot.
2. The specifics of facility research which is naturally multidisciplinary presents both challenges and opportunities for the promotion of new PID types and the PID graph. The main challenge is different practices of research communities involved with facilities, specifically the practices of publishing research results and the practices of acknowledging facilities' support for research. This diversity is also an opportunity, as research communities with better practices can serve as role models for others.
3. The Portal development will be sustained using STFC's own resources, for which internal funding is secured until the end of the financial year 2020/21 (March 2021). This will allow progress with the technological aspects of the Portal, also with widening participation of various STFC stakeholders in the Portal's evaluation and requirements gathering.

[47] DAMDID 2020. 13-16 October 2020. http://damdid2020.cs.vsu.ru/conference_program.html
[48] Mini-ELAG 2020. 20 October 2020. https://elag.org/2020/09/24/mini-elag-program/

3 Jupyter notebooks illustrating the PID Graph

3.1 Background

The FREYA project developed a set of user stories representing a wide range of opportunities for the PID Graph benefitting different stakeholders. These user stories, which are available on GitHub[49], have been used to motivate the development of the project's own pilot applications and the technical bases of the PID Graph.

A subset of these user stories have been further developed through a subcontract in the form of Jupyter notebooks which use FREYA services and tools, i.e. the DataCite GraphQL API, to construct specific PID graphs and present them in a clearly understandable format to incentivize users to interact with the PID Graph and discover content. Jupyter notebooks easily integrate with the standardized GraphQL query interface and provide an easy to use platform for addressing PID Graph user stories.

This chapter presents the description of work for the subcontract for the creation of the Jupyter notebooks, the website that was developed as an alternative setup to showcase the notebooks, and the outreach activities that were carried out to promote this work as well as to collect feedback.

3.2 Subcontract work

FREYA partner STFC paid a subcontractor (as foreseen in the FREYA Grant Agreement) to develop and document a number of Jupyter notebooks from May to August 2020, which address specific FREYA user stories. FREYA partners identified more than 40 user stories in 2018 which were initially documented via GitHub, and later in the PID Graph section of the PID Forum[50].

For the purposes of this work, 10 of those user stories were chosen, which address user stories from various disciplines and different stakeholder groups, which have been tested and fully documented. Each user story was developed into a notebook and all 10 computational notebooks are hosted on a GitHub repository[51].

They are made available with an MIT Open Source license and it is possible to run them easily via Binder from GitHub. After completion of the work on these computational notebooks, they were all registered using codemeta metadata and a DataCite DOI – this also makes them available in the PID Graph.

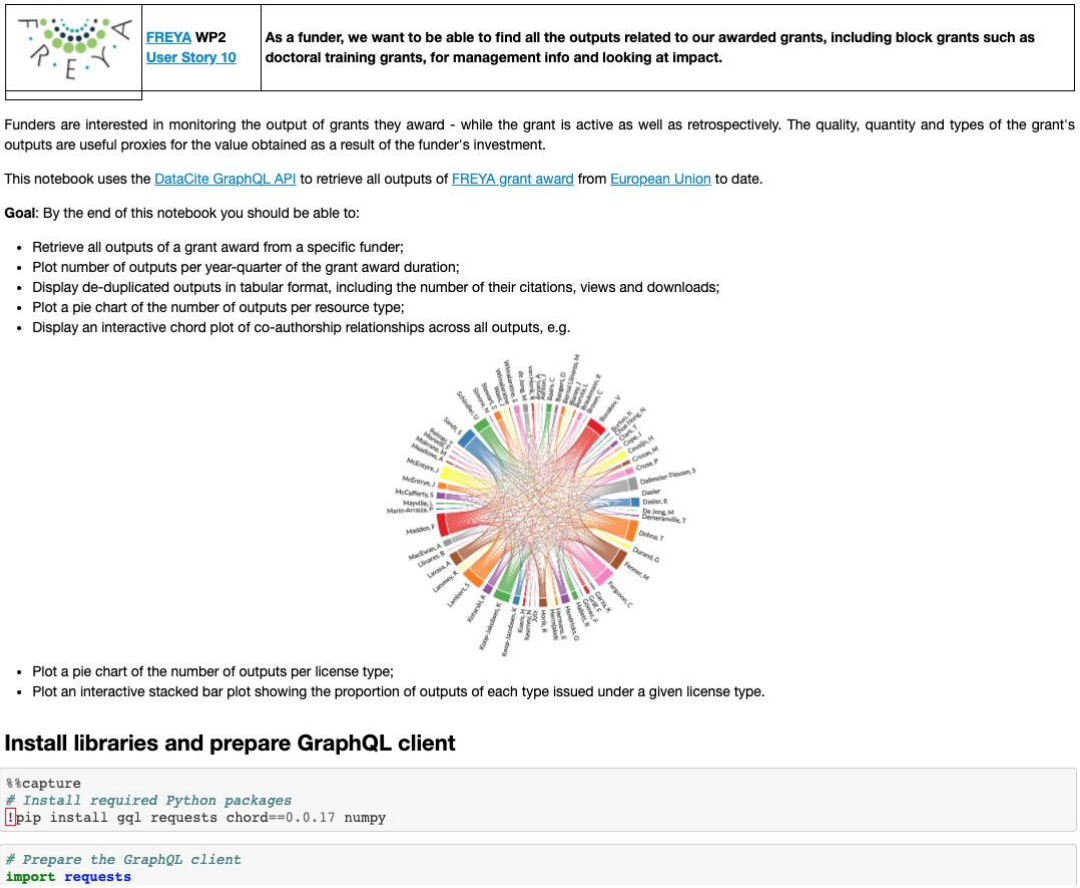
The specific user stories that were developed are shown below in Table 1, including the PID entities which are applicable for each case, and the DOI for the computational notebook source code:

[49] FREYA PID Graph user stories on GitHub: https://github.com/datacite/freya/issues?q=is%3Aopen+is%3Aissue+label%3A%22PID+Graph%22
[50] FREYA PID Graph user stories on the PID Forum: https://www.pidforum.org/c/pid-graph/17
[51] FREYA Jupyter notebooks GitHub repository: https://github.com/datacite/pidgraph-notebooks-python

User story number	PID entities	User story description
1	Dataset, Publication	As a data center, I want to see the citations of publications that use my repository for the underlying data, so that I can demonstrate the impact of our repository. DOI: https://doi.org/10.14454/r0ed-fh20
2	Software, Researcher	As a software author, I want to be able to see the citations of my software aggregated across all versions, so that I see a complete picture of reuse. DOI: https://doi.org/10.14454/27b7-9g84
3	Research organization, Publication, Dataset	As an administrator for the University of Oxford I am interested in the reuse of research outputs from our university, so that I can help identify the most interesting research outputs. DOI: https://doi.org/10.14454/e23v-x328
4	Funder, Publication, Dataset, Software	As a funder I want to see how many of the research outputs funded by me have an open license enabling reuse, so that I am sure I properly support Open Science. DOI: https://doi.org/10.14454/q7jn-xw50
5	Publication, Research organization	As a student using the British Library's EThOS database, I want to be able to find all dissertations on a given topic. DOI: https://doi.org/10.14454/jkar-xj80
6	Researcher	As a researcher, I am looking for more information about another researcher with a common name, but don't know his/her ORCID iD. DOI: https://doi.org/10.14454/03vp-ry06
7	Publication, Dataset	As a data center, I want to see the citations of publications that use my repository for the underlying data, so that I can demonstrate the impact of our repository (second degree citations). DOI: https://doi.org/10.14454/xw22-0w50
8	Dataset	As a longitudinal study, I want to be able to deduplicate the metrics/impact for our data, so that I can see the impact of our study's data as a whole. DOI: https://doi.org/10.14454/y785-xs19
9	Researcher, Publication, Dataset, Software	As a bibliometrician, I want to know all the co-authors of a particular researcher, so that I can do a network analysis of the researcher's collaborations. DOI: https://doi.org/10.14454/62t3-0822
10	Funder, Publication, Dataset, Software	As a funder, we want to be able to find all the outputs related to our awarded grants, including block grants such as doctoral training grants, for management info and looking at impact. DOI: https://doi.org/10.14454/qaym-kt26

Table 1: The 10 Jupyter notebooks developed as part of the subcontract including their DOIs and a mapping of PID Graph user stories to PID entities.

All notebooks are written using the JupyterLab software[52] and using the language Python in version 3.6. The readme file of each notebook contains a link to run the computational notebook in Binder software[53]. All notebooks contain detailed inline documentation, e.g. notebook 10 in Figure 21 below:



FREYA WP2 User Story 10

As a funder, we want to be able to find all the outputs related to our awarded grants, including block grants such as doctoral training grants, for management info and looking at impact.

Funders are interested in monitoring the output of grants they award - while the grant is active as well as retrospectively. The quality, quantity and types of the grant's outputs are useful proxies for the value obtained as a result of the funder's investment.

This notebook uses the DataCite GraphQL API to retrieve all outputs of FREYA grant award from European Union to date.

Goal: By the end of this notebook you should be able to:

- Retrieve all outputs of a grant award from a specific funder;
- Plot number of outputs per year-quarter of the grant award duration;
- Display de-duplicated outputs in tabular format, including the number of their citations, views and downloads;
- Plot a pie chart of the number of outputs per resource type;
- Display an interactive chord plot of co-authorship relationships across all outputs, e.g.

- Plot a pie chart of the number of outputs per license type;
- Plot an interactive stacked bar plot showing the proportion of outputs of each type issued under a given license type.

Install libraries and prepare GraphQL client

```
In [29]: %%capture
         # Install required Python packages
         !pip install gql requests chord==0.0.17 numpy

In [30]: # Prepare the GraphQL client
         import requests
```

Figure 21: Inline documentation of computational notebooks, here notebook 10.

The notebook output is typically one or more visualizations, tabular data and/or data in other formats, e.g. bibtex. One notebook (notebook 9) generates data that can then be imported into the VOSviewer open source tool (Figure 22)[54]:

[52] JupyterLab: https://jupyterlab.readthedocs.io/en/stable/

[53] Binder: https://mybinder.org/

[54] Perianes-Rodriguez A., Waltman L., van Eck N. J. (2016): Constructing bibliometric networks: A comparison between full and fractional counting. Journal of Informetrics, 10(4), 1178–1195. https://doi.org/10.1016/j.joi.2016.10.006

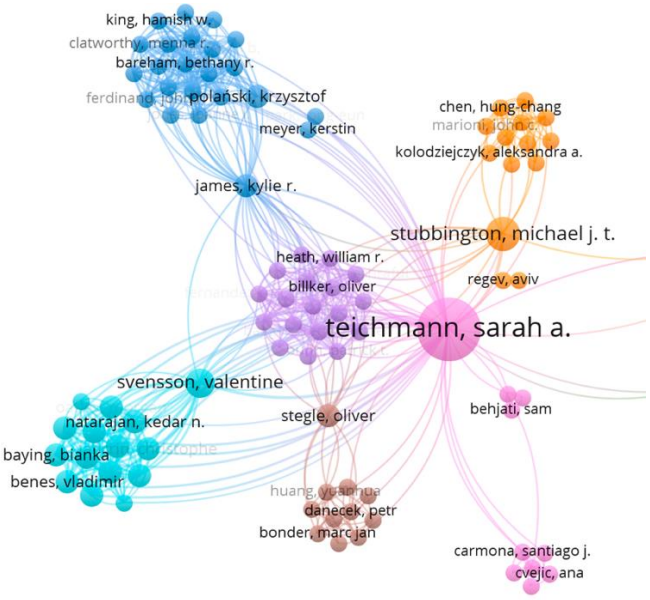


Figure 22: Visualization of co-author networks using notebook 9 and the VOSviewer tool.

Since the main purpose for the notebooks is to incentivize various communities to interact with the PID Graph and discover its content, while choosing which use cases to bring forward as a computational notebook it was important to consider how the FREYA partners relate to each use case that was selected. For that purpose, the disciplinary partners were asked to indicate which of the user stories (and resulting notebooks) they considered relevant for their organization/community (Table 2).

User story number	FREYA partners
1	CERN, British Library, DANS, PANGAEA
2	CERN, STFC
3	PANGAEA
4	EMBL-EBI, STFC, DANS, PANGAEA
5	British Library
6	CERN
7	CERN, EMBL-EBI, DANS, PANGAEA
8	British Library, PANGAEA
9	British Library, EMBL-EBI
10	EMBL-EBI, STFC, DANS, PANGAEA

Table 2: Mapping of PID Graph user stories to disciplinary partners.

From this exercise we were able to ascertain that the selected notebooks could possibly have some application to the disciplinary communities represented in FREYA. The final intention for the notebooks is for external people to be able to pick them up and adapt them for their own use - that and the website for

showcasing them (see next chapter) are offerings from FREYA that allow for easier discoverability of PID Graph assets.

In summary, the FREYA project has created 10 computational notebooks using the popular Jupyter framework, addressing important user stories identified in the FREYA project, and made them freely available with an open source license. The notebooks can be easily reused by using different PIDs and/or queries as input parameters.

3.3 PID notebooks website

In parallel to the work carried out by the subcontract, we started working on launching a website to showcase the notebooks[55]. The idea behind this was to provide a static website as an alternative way of accessing the notebooks, which is faster and easier than having to run the nbviewer server in order to render the notebooks (Figures 23 and 24).

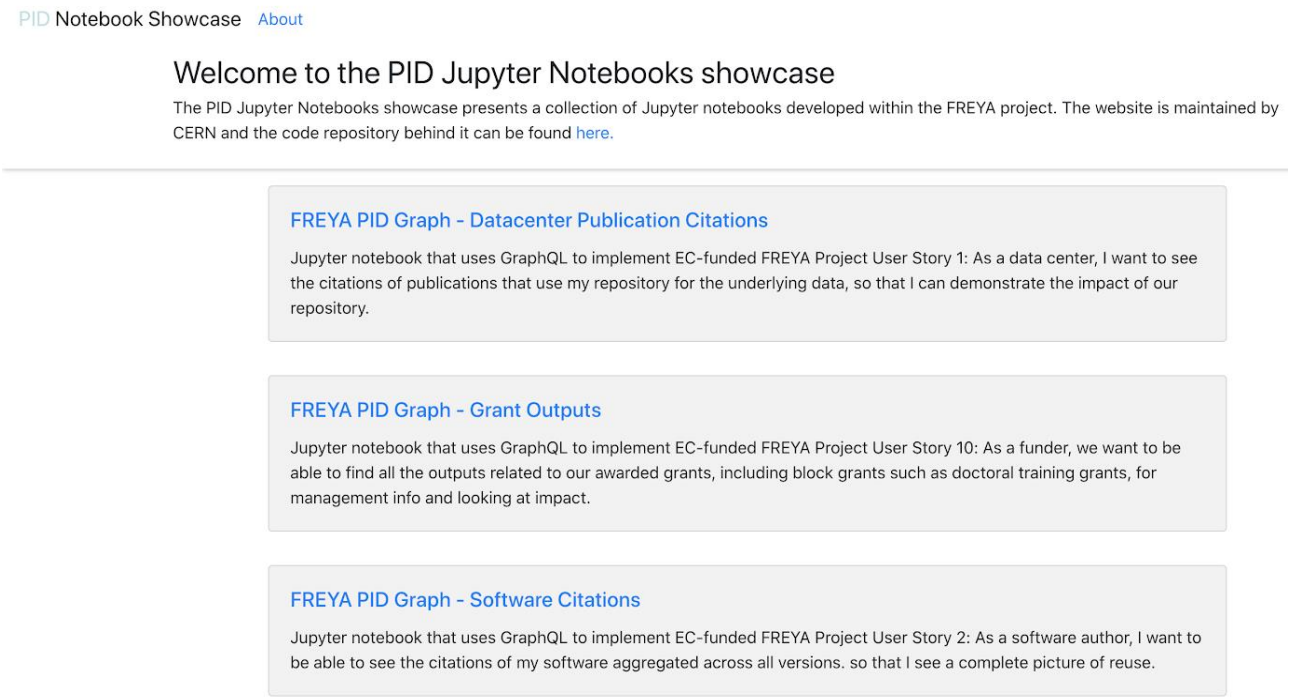


Figure 23: pidnotebooks.org site user interface for the welcome page.

[55] PID Jupyter notebooks website: https://pidnotebooks.org/

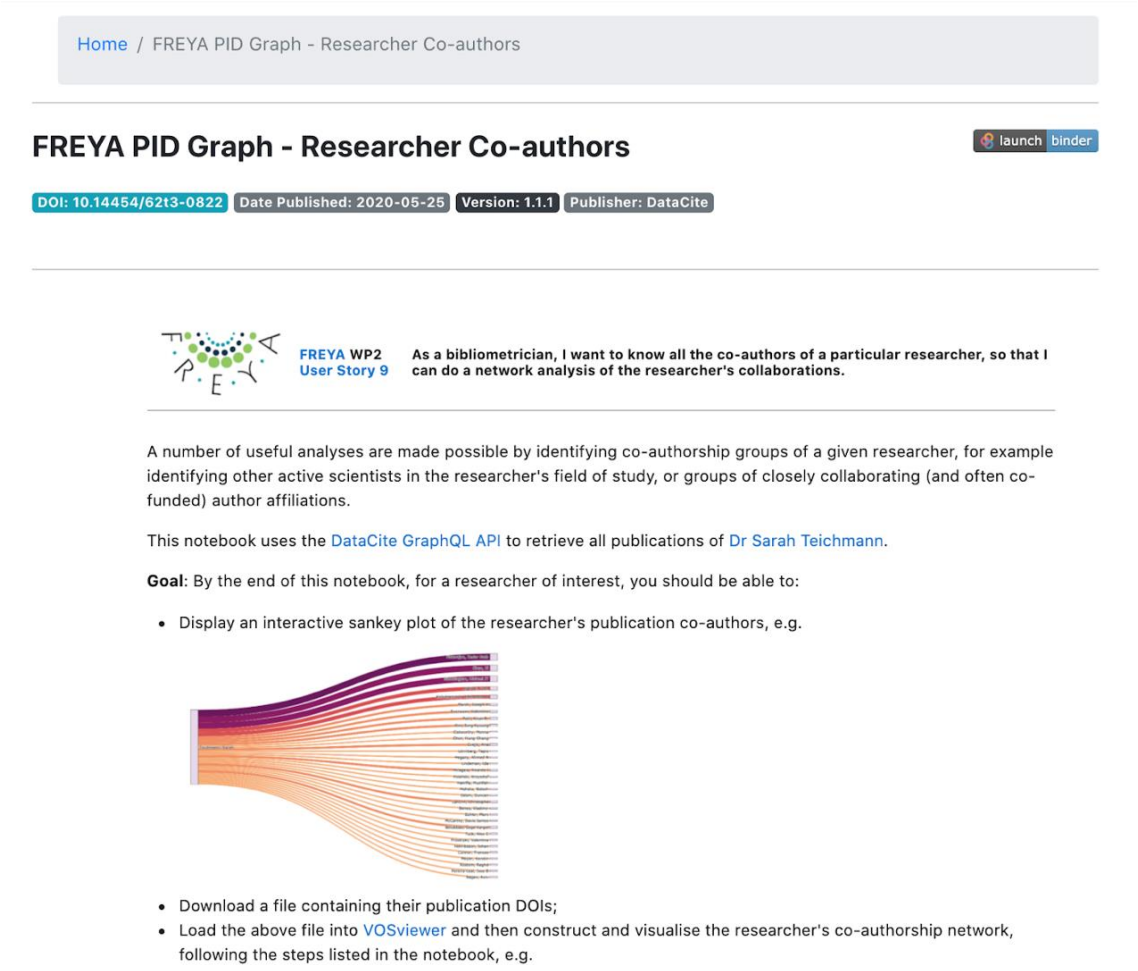


Figure 24: pidnotebooks.org detailed record page for notebook 9.

As CERN decided to do the work on the development of the website, a CERN GitHub repository was used to host the code[56]. The site will be maintained by CERN and the domain by DataCite. It is not expected that it will be updated though. The code for the website is open source and available under the MIT license.

In terms of styling, the website was made to look as much as possible like the PID Services Registry website[57] which had been launched earlier to allow for more consistency across FREYA products. The choice behind the domain name followed the same logic, i.e. “pidnotebooks.org” to be similar to “pidservices.org”. The appropriate acknowledgements are stated at the bottom on the website (FREYA logo, EC funding).

The website is created using Gatsby[58], a static website generator. Gatsby uses the power of technologies like GraphQL and React, to generate SEO ready (search engine optimization) websites with minimal configurations. With continuous integration tools, every time a change is committed in the code, an updated website is generated and deployed. The python notebooks to be displayed, are placed in a “contents” directory and automatically fetched by the CI scripts to generate the new HTML that will be shown. For each individual notebook, information is also fetched from the "codemeta.json" files which contain (software) metadata and were generated for each notebook as part of the subcontract work[59]. The

[56] GitHub repository for the FREYA Jupyter notebooks website: https://github.com/cernanalysispreservation/freya-pid-notebooks-showcase
[57] FREYA PID Services registry: https://pidservices.org/
[58] Gatsby: https://www.gatsbyjs.com/
[59] Example of a codemeta file for one of the FREYA PID notebooks: https://github.com/datacite/pidgraph-notebooks-python/blob/master/user-story-10-grant-outputs/codemeta.json

DOI and Binder links are also included in each record which makes it possible to launch a notebook in a Binder instance directly from the website.

The website was released in October 2020 (TRL9), and minor updates and bug fixes may continue until the end of the project in November 2020.

3.4 Outreach and feedback from the community

The online PID Forum (pidforum.org) played a central role in our outreach activities for the Jupyter notebooks. The use cases and GraphQL queries that we developed were promoted within the broader PID community through dedicated sections on the PID Forum[60]. We also established a topic on the use of Jupyter notebooks and GraphQL to provide more background and link to relevant information and materials[61].

To showcase the notebooks to a wider audience and to provide the opportunity to gather feedback directly, FREYA organized a dedicated webinar on the 27th of August 2020[62]. In this webinar, we first gave a brief introduction on the PID Graph and its concept and then focused on the user stories and their implementation through the Jupyter notebooks. At the end of the webinar there was room for questions from the audience which resulted in an active discussion on the notebooks and its use for the webinar attendees. There were 31 attendees at the webinar itself, and the recording that was made available on the FREYA YouTube channel[63] has thus far already attracted around 90 views (as of mid-October 2020), indicating a good interest from the community in our notebooks.

After the webinar - which marked the initial release of the notebooks - FREYA asked for community feedback through a survey. The survey was promoted to the general PID audience on the PID Forum, as well as specifically sent to the FREYA ambassadors, the webinar participants and representatives from several EOSC Cluster projects that participated in a FREYA workshop on the role of PIDs in disciplinary systems.

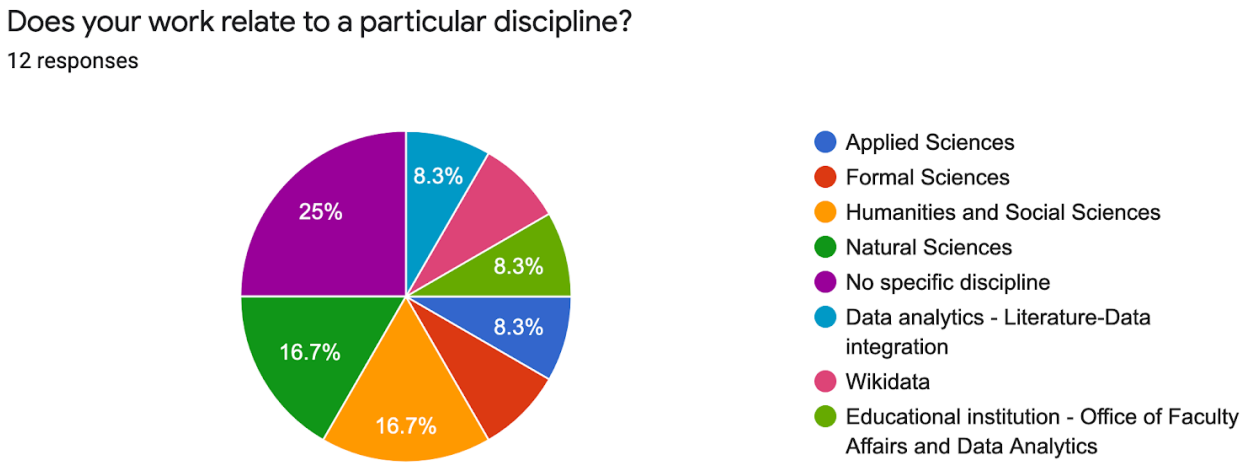


Figure 25: Disciplinary breakdown of survey respondents.

[60] Example of a section on the PID Forum discussing the Jupyter notebook and its related use case on software https://www.pidforum.org/t/pid-graph-graphql-example-software-versions/932

[61] PID Forum section on Jupyter notebooks and GraphQL: https://www.pidforum.org/t/using-jupyter-notebooks-with-graphql-and-the-pid-graph/392

[62] FREYA webinar on Jupyter notebooks -materials: https://zenodo.org/record/4004426#.X3syPZMzYUE

[63] Recording of the webinar on the FREYA Youtube channel: https://www.youtube.com/watch?v=I7MUTFvjPzo&t=50s

The respondents (n=12) were from various disciplines and professions (Figure 25), with researchers (33%) and developers (25%) being represented most. Their relationship with FREYA was just as diverse; other than FREYA ambassadors (50%), the audience were also webinar attendees (42%) or PID enthusiasts (58%)[64].

Interestingly, respondents indicated that they were not very familiar with the notebooks (58% claimed they were not at all familiar or not very familiar with them), and all but one said that they had not yet used the notebooks to query the PID Graph. However, participants did report the notebooks to be useful (see Figure 26 below) and 42% indicated that they were planning to use the notebooks in the future.

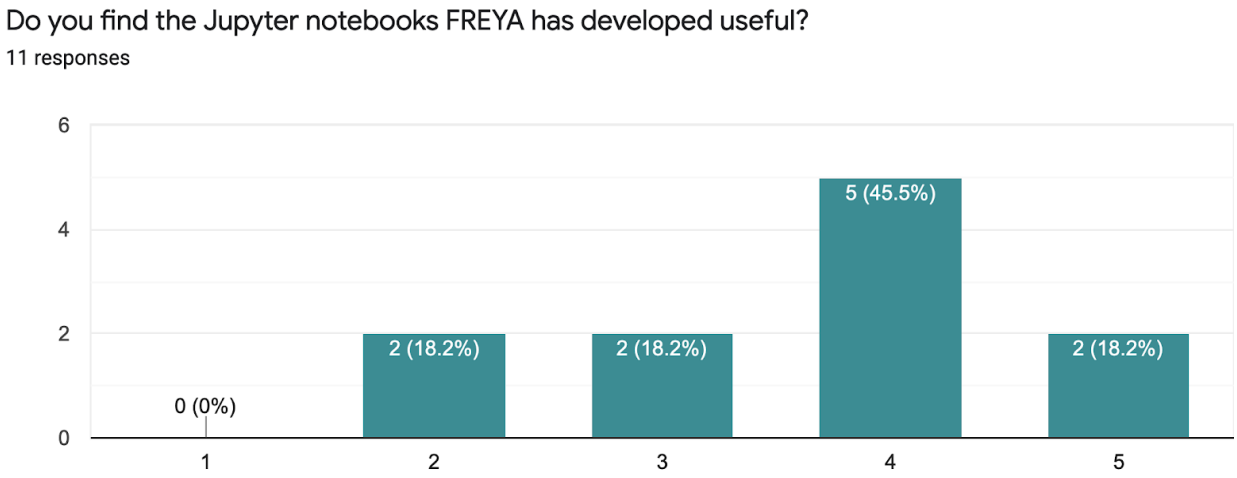


Figure 26: Responses from the questionnaire on the FREYA notebooks. The number 1 corresponds to "not at all" while number 5 corresponds "very much".

We asked participants which notebooks they found particularly interesting for their own work and the two notebooks that were mentioned most often were notebook 9 on network analyses and research collaborations mentioned by 6 respondents, and notebook 3 on research outputs from universities mentioned by 5 respondents (see Chapter 3.2 for more details on individual notebooks). Since most respondents had not yet used the notebooks, we acquired little feedback on issues or suggestions for changes to the notebooks from this survey. Several respondents indicated that they needed to become more familiar with the notebooks and python programming before being able to provide additional feedback.

The last months of the FREYA project were used to further promote the notebooks at our events, including for instance the FREYA final event[65], and through the Notebooks website described in the previous chapter (3.3).

[64] Multiple choice question.

[65] The FREYA Final event: Realising the European Open Science Cloud will be held in November 2020 and includes a presentation of various FREYA PID services including the Jupyter notebooks: https://www.project-freya.eu/en/events/joint-eosc-hub-freya-sshoc-event

4 Lessons learned and moving forward

This deliverable presents the final form of the FREYA pilot applications and the progress made within the timeframe of the project. Some of the pilot applications included in the work of WP4 were services that already existed or even well-established services, e.g. the British Library's Explore, NARCIS, Europe PMC, etc., which used PIDs to some degree before FREYA. For such services, work during the project was focused on enhancing PID Graph aspects for which there was already a basis - adding more support for mature PIDs and integration of new/emerging PID types, for example, or better interlinking of PIDs. At the same time, entirely new pilots were also developed during the project to address various user stories or needs in a user community which will be taken forward after the project ends - examples of that are PANGAEA's IGSN work or STFC's Open Science Portal.

The pilot applications demonstrate how the concept of the PID Graph and the work developed by core PID service providers in WP2 can be implemented in community services and how communities can be incentivized to discover and exploit the content of the PID Graph. As far as specific integrations into WP4 pilots of tools or services that were developed in WP2, various partners have stated working concretely on that front - e.g. support for the GraphQL API by DANS, STFC, etc. or provenance tracking in PANGAEA's case. The Jupyter notebooks developed during the project are a separate mechanism for demonstrating the capabilities of the PID Graph and an important - and highly adaptable - PID discoverability tool (facilitating discovery by various communities in general, discovery of new PID types or interconnections, etc.). While the survey on the notebooks yielded only a small amount of responses given the limited time available, it also provided a first glimpse into the attitudes towards the idea of the notebooks as a discoverability and exploitation tool for the PID Graph, which was positive.

As with all WP4 deliverables, what is evident here as well is that each community faces different challenges and has distinct approaches and vision for PID Graph development in the future. This is primarily reflected in what each disciplinary partner decided to work on within the context of the FREYA Grant Agreement and the variety of user stories that were collected which address the needs of each community. Given the range of disciplines represented in the project, that is to be expected, but at the same time it is possible to find commonalities especially when it comes to observations for the future and lessons learned.

PIDs are as diverse as the landscape of research practices. Every community and every discipline has their own characteristic research practices, and interacts with specific and maybe unique research objects. This makes the choice of relevant PID types community-specific, as pointed out by the British Library and EMBL-EBI for example. This multidisciplinarity and community-specific behaviors proved to be a challenge for the integration of new PID types, as there is not one "correct" or common way to serve all. The diversity of approaches and practices complicated, for example, the normalization of metadata elements from different resources, as metadata standards vary and PID standards for different research objects are not always on the same maturity level. However, this great diversity may also be a benefit, as disciplines that already adopted a variety of PIDs into their workflows could serve as precedent for others and encourage the use of new PID types.

From the more "technical" side of things, as stated by many partners, the complexity of a rapidly-evolving PID landscape became challenging for users that are not familiar with PIDs - for example the fact that there are many different PID types for a single object is at times confusing and limiting. In some cases (e.g. British Library), decreasing the amount of PIDs displayed in the frontend of a service contributed towards a better user experience. Not only PIDs, but also APIs proved to be diverse, which made it difficult to connect different services within a PID graph. What was identified as challenging in particular, was the ability to connect specific endpoints where the objects were dynamic, which impacted the accuracy of links between PIDs.

What became clear is that there is a need for incentives for the integration of new PIDs in existing services, as service managers were hesitant to prioritize the integration of features that are not mature yet. A similar

need was identified for users, as they need to invest extra time and effort into providing machine-readable metadata for external PIDs without having a clearly visible benefit from it; this is an issue that needs to be addressed possibly through outreach and training materials from the side of the service providers, and by making it as easy as possible for users to use PID-related functionality. What's more, user support by curators during the data submission process, and dedicated publication workflows that automatically generate PID-related metadata may assist in solving this issue.

The founding of an overall PID agency, which stimulates the use of PIDs across multiple communities, was identified as a potential future solution. Partners noted that such an agency could be responsible for a commonly-accepted core set of minimal PID standards, and generally help to resolve those community-specific issues which general PID guidance sometimes fails to reconcile. Work on this topic has been introduced by the FREYA Sustainability Work Package (WP6).

Many partners considered the PID Forum, the PID Graph visualizations, and the Jupyter notebooks as meaningful tools that solve many of the identified issues, as they broaden the reach of crosslinking, reduce barriers to adoption of new PIDs, and serve as incentives for the users to provide qualitative metadata. Drawing on these resources, and adding new user stories to the PID Graph in the future is something that FREYA partners are interested in.

Besides these more general lessons learned and future plans, some partners brought up issues that are specific to their community, or concerned only specific PID types which are worth highlighting:

- The British Library identified the need to improve the integration between ARKs and its library system in order to make ARKs externally resolvable.
- CERN identified the need of a single entry point which in the background connects all the local graphs of each service so that users can fetch information from all those services through a central place.
- DANS emphasized the need of PIDs for all sorts of research projects and noted that it will require more outreach work to show the value of PIDs for projects to make progress.
- PANGAEA addressed the need of a better integration of formats for IGSNs, as well as machine-readability of metadata content, and allowing for easier discoverability of samples through associated research. The scalability of services and technical infrastructure will be a great challenge for broader use of IGSNs in the future.

The services presented in this deliverable will continue to be developed, and there is potential for further integration into EOSC beyond what has been accomplished in the context of the FREYA PID Services Registry. The work introduced in D4.5 about integrating the PID Graph with EOSC underlined, from the point of view of other projects (e.g. EOSC cluster projects) within the disciplinary communities represented in FREYA, use cases and potential regarding PID functionality/PID services in EOSC. All in all, the pilot applications in FREYA have progressed significantly since D4.1; the disciplinary partners will continue to use the tools and practices developed throughout the lifetime of the project to further build on their services for the benefit of their communities.