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

D4.3 Using Advanced PID Graph Functionality in Pilot Applications

Deliverable type	Report
Dissemination level	Public
Due date	31 May 2019
Authors	Tina Dohna (PANGAEA, orcid.org/0000-0002-5948-0980) Ketil Koop-Jakobsen (PANGAEA, orcid.org/0000-0002-1540-6594) Uwe Schindler (PANGAEA, orcid.org/0000-0002-1900-4162) Artemis Lavasa (CERN, orcid.org/0000-0001-5633-2459) Pamfilos Fokianos (CERN, orcid.org/0000-0003-0618-1722) Sünje Dallmeier-Tiessen (CERN, orcid.org/0000-0002-6137-2348) Vasily Bunakov (STFC, orcid.org/0000-0003-3467-5690) Frances Madden (British Library,orcid.org/0000-0002-5432-6116) Christine Ferguson (EMBL-EBI, orcid.org/0000-0002-9317-6819) Chris Baars (DANS, https://orcid.org/0000-0002-5228-1970)
Abstract	This deliverable describes the work carried out on integrating PID Graph functionality in the disciplinary contexts of FREYA consortium partner systems, demonstrating the diversity of application of the PID Graph idea.
Status	Submitted to EC 21 June 2019 Revised version submitted to EC 8 November 2019 *Summary table added per pilot application to improve the comparability among pilot applications. Addition of description of pilot application of project partner DANS. Applicability of the results to EOSC emphasized and summarized.*

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 organizations 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

The PID Graph, in the context of the FREYA project, is conceived as a federated network of thematic graphs, which are implemented by project partners and built along a common strategy and similar technology to allow the graph to grow in a coordinated approach. Advancing the graphs' functionality in different disciplinary contexts was at the core of the work, which is now being described in this deliverable. Each participating partner (PANGAEA, BL, CERN, EMBL-EBI and STFC) describes pilot applications that they conceived and designed for users, who want to benefit from graphed PID connectivity of related research items. Additionally, DANS describes their pilot application, which was developed in addition to task 4.5 of the FREYA project.

The tools and services were developed to accommodate user needs, either following relevant user stories collected by Work Package 3 to focus requirements for selected new PID services (see D3.2[1]) or along user needs arising from the operational activities of the partner. Altogether, this deliverable describes five use cases facilitating an advancement of the existing PID graph functionality by FREYA partners. A diverse set of "PID graph" navigation-tools were built around a wide variety of PIDs:

- PANGAEA focused on providing a tool for navigating a PID graph revolving around PIDs for natural samples (IGSNs) proving the user with metadata and links to related samples, datasets and publications etc.
- British Library provided tools that consolidate identical documents or identify different versions of the same document in a format that can be easily understood by users.
- EBI provided a tool to manage duplication and versioning of preprints.
- CERN provided a functionality for fetching metadata from external sources in the CERN-community PID graph, and a visualization tool for Python Notebooks to illustrate direct data access in the PID Graph.
- STFC provided a tool to track the impact and activities of PhD student using STFC facilities.

Each application provides background information of the disciplinary context in which it was developed and of the partners addressing it, followed by a detailed description of the application, the current state of technological development. All tools and services are expected to be realized within the time frame of the FREYA project. Currently, they are at different development stages, ranging from blueprints to full technological development and ready for implementation.

A primary focus for FREYA is to act as a catalyst for the maturation and adoption of new PIDs for research entities, so that new PID nodes populate the landscape and can be incorporated in graphed networks. Pilot applications, like the ones described here, demonstrate the added value of PID implementation and can create the needed momentum for more large scale and across the board PID use by all actors in the data life cycle. If the utility of PID and PID Graph services proliferates in the research communities then PID adoption and maturation of new PIDs will accelerate, creating the needed momentum.

Each of the pilot applications provides a service for the usage of PIDs to explore linked data and knowledge allowing research outcome in the form of data and/or other digital objects to be easily findable, accessible, interoperable and reusable (FAIR). PIDs are envisaged to play a primary role in EOSC, and once these services have reached the maturity level required by EOSC, they can be implemented as part of the EOSC services registry.

Summary tables are generated for each pilot applications allowing for direct comparison of PID-graph functionality, methods, rationale, target-users and EOSC Applicability.

[1] https://www.project-freya.eu/en/deliverables/freya_d3-2.pdf

Contents

1 Introduction

One of the aims of FREYA is to set up an infrastructure for Persistent Identifiers (PIDs) to support and enable Open Science, providing core services across Europe and globally. Generating a user-based infrastructure that will ease findability, accessibility, interoperability and reuse is essential for progression and implementation of Open Science by the research communities. While populating and maturing the PID landscape is an essential precursor, this deliverable reports on developments by consortium partners that enable users to navigate the PID landscape to satisfy specific information needs.

1.1 PID implementation

Improved implementation and usage of PIDs by the relevant communities is a key-task in FREYA. PIDs for key entities in science, such as researchers (ORCID), publications (DOI), and data (DOI), have already been implemented and have reached a maturity level where PID assignment and large-scale application is well-established. Significant improvements can be facilitated by an expansion of this PID landscape through implementing new and emerging PID systems (e.g. software, samples and organizations; see FREYA D3.1[2]) in order to increase the amount of accessible information. This will facilitate and improve access to the metadata available for any given entity and provide options to enrich metadata with linked entities and resources. For example, implementation of IGSNs (International Geo Sample Number - a PID for physical samples) provides access to a standard set of important sample parameters through the metadata available from the IGSN landing pages. Including additional PIDs within this metadata can allow, for example, the identification of all the researchers who have worked on the sample throughout the years and their scientific output in form of published data and scientific articles by citing the sample ID in their research. Another example is the implementation PIDs for software, where metadata about the software becomes available through the software PID landing page. Most software packages include regular updates and changes, so that detailed information is necessary to replicate analyses accurately. Furthermore, linking software PIDs to other PIDs can allow users to easily discover published data derived from the same or similar versions of the software, discover who has used the software and published output from the analyses.

1.2 PID Graph functionality

Paving the way for the implementation of new PIDs is a key task in FREYA. Implementation of new PID systems will in itself provide access to essential metadata in standardized formats and facilitate the linkage of related PIDs and the entities underlying them. However, it still requires the user (the researcher using the PIDs to search to information) to actively browse through various online repositories and databases in order to acquire the desired information. Specific skills and experience are often required to efficiently navigate the PID landscape to gather all the related resources that the user may be interested in. Missing here are tools that users can apply to navigate this network of interconnected PIDs (PID Graph) and allow easy and reliable access to the needed pool of resources. Enabling users with these tools is what we refer to as PID Graph functionality, which the disciplinary partners are implementing in their systems to empower users to use PIDs to their fullest potential.

Consequently, facilitating PID implementation in itself is not enough and services are needed to help the users to navigate the PID Graph, thereby increasing the functionality of the PID Graph. The value of the PID infrastructure can be significantly boosted through the development of services increasing PID Graph functionality and helping users to navigate the Open Science landscape. Ideally, tools should presuppose that the users do not possess any prior knowledge of PIDs and their functionality. The created services should provide a user-friendly interface, granting users an entry point to the PID Graph

[2] https://www.project-freya.eu/en/deliverables/freya_d3-1.pdf

and simple functions to navigate the network and Open Science landscape. Prior to work done by FREYA partners, the Technology Readiness Level (TRL) of the PID landscape supported this kind of functionality only for the three most mature PID entities: ORCID for researcher and DOIs for articles and data. Despite the maturity of the these PIDs, even here we can find the functionality sometimes limited due to, for example, missing backward implementation in older studies.

1.3 PID Graph functionality in FREYA

In FREYA, services for increased PID Graph functionally take a starting point in the expansion of the amount of functional PIDs linking emerging PIDs to existing and mature PIDs. Such services, for example, include improved data citation or software citation workflows, as well as improvements in data searching functionality through both disciplinary data discovery systems and in the core PID systems. The individual FREYA partners are working on improving PID Graph functionally within their domains. In this deliverable, we explore how FREYA partners have advanced their PID Graph functionality by building tools or optimizing workflows helping users to navigate the individual PID Graphs or specific parts thereof. Use cases are employed to exemplify specific user-needs and demonstrate how improved PID Graph functionality can improve the user's outcome of exploring the PID Graph.

It is the aim of FREYA to make their developed services widely available within the context of Open Science research. As part of the work on increasing the PID Graph functionality, individual partners develop specific tools and services that work within their local PID Graph, but may also be useful in a wider context e.g. by data-repositories and data managers from other disciplines. In order to facilitate a wider usage, it is the goal to have as many of the tools and services as possible registered in the EOSC services registry. However, the EOSC service catalogue is restricted to services with TRL 8 and 9[3] – these are the services considered mature enough to be at production level. Hence, it will be difficult for technological development initiated within the FREYA to reach this level within the timeframe of the project.

[3] https://www.eosc-hub.eu/news/eosc-hub-contribution-eosc-open-consultation-rules-participation

2 Methods

The scope of this deliverable is to exploit advanced PID Graph functionality in disciplinary contexts in accordance with Task 4.5 of the FREYA work plan.

Advances in PID Graph functionality implemented by individual FREYA partners are demonstrated in various use cases. Furthermore, advances in PID Graph functionality developed by WP2 (PID core services) will be integrated by partners and demonstrated in this deliverable-report.

Individual FREYA partners have integrated methods facilitating direct access to content that is linked to PID landing pages of disciplinary systems in accordance with Task 4.5.1 of the FREYA project, and furthermore explore the advantages of PID resolution of the same resource at multiple locations in accordance with Task 4.5.2.

2.1 Use cases as a methodological approach

In WP4 of the FREYA project, the focus is on users from various research communities. To ensure the advancement of PID graph functionality specific use cases acted as starting points. Technological advancements can be achieved at many levels and by choosing use cases as a key component of our advancements in PID graph functionality, we want to ensure a user-centered point of view that guarantees developments that actually benefits the end-users, and facilitates uptake of PIDs in the communities. In the context of FREYA, a use case revolves around users of Open Science, and describes an actual scenario, where technological developments can fulfill a specific user-need, and thereby improve the user experience. FREYA has previously collected a list of more than 75 use cases from the Open Science community[4]. These use cases provided the foundation for identifying domain specific user needs. These needs can be accommodated through the linkage of PIDs in networks (a PID Graph). In task 4.5, FREYA-partners then developed the necessary tools to navigate this PID graph, allowing the users to exploit the enriched amount of information, which can be retrieved through the combined access to related PIDs. All partners in FREYA have infrastructures reliant on PIDs and tools to access the PID associated metadata. These PID Graphs, for the most part, accommodate the basic user-needs in terms of retrieving metadata and providing links to related information. However, the current PID Graph functionality does not accommodate all user-needs and the graphs are continuously developed and expanded. Increasing the PID Graph functionality is essential to generate a better user-experience.

In this deliverable, individual FREYA-partners from PANGAEA, British Library, CERN, EMBL-EBI and STFC, demonstrate how they can accommodate user-needs through advancement of their PID Graphs. Each partner presents a specific use case and demonstrates how an advancement of their PID Graph functionality can facilitate improved discoverability, better organization and improved visualization of PID related metadata. To facilitate increased PID graph functionality, each partner has planned and/or developed prototype applications to advance the exploitation of their PID graph.

The prototype applications are at different stages of technological development, and this deliverable focuses on the conceptual developments behind the improved PID Graph functionality. It identifies new PID Graph functionality that will fulfill specific needs for the PID Graph use, and provides a blueprint for the development that is either needed or already conducted by the individual partners. A common denominator for the presented use cases is that the work on the PID Graph within FREYA has facilitated these PID Graph advancements.

[4] https://www.pidforum.org/c/user-stories

2.2 Comparative presentation of the established advancement in PID Graph functionality

The following FREYA-partners have advanced their PID Graph functionality in connection with this deliverable:

PANGAEA has implemented actionable links to IGSNs (PIDs for physical samples) in their data repository, resolving directly to the landing pages of the specific IGSN. Furthermore, PANGAEA increased PID Graph functionality by designing an app (FREYA IGSN Scanner), which can extract essential metadata information through the linkages of IGSNs, ORCIDs, and DOIs for articles and data, and make it available in an easily accessible format for users.

The British Library will implement improvements to the discovery and access to its resources, especially articles, across multiple locations by implementing an API and microservice which will use an article's DOI to allow access options for resources to be displayed in one place, improving end users' experience. This API could also be applied by other organizations trying to streamline access across different locations.

CERN has enhanced "*CERN Analysis Preservation*"; a community PID Graph to preserve and identify research objects, by improving the automatic population of records with information from external sources, which includes better ORCID integration, and by creating a visualization tool that integrates Python notebooks.

At EMBL-EBI, the literature repository Europe PMC is indexing preprints, and thereby expanding its scope of literary content. Preprints are considered to be precursors to full articles that are accepted by publishers after peer review. They can also be revised prior to acceptance and be indexed as different versions, each of which contains demonstrably different information available for citation and linking to ORCID profiles. Reporting for this deliverable, Europe PMC is working to expose the version information associated with preprint records, to enrich the provenance information available to users. This information is now available for programmatic access via the Restful API and will soon be available in the user interface. Differentiating versions is a common challenge in scholarly research and the approach taken here for literature could be applied to other research outputs, across disciplines.

STFC has worked out a user story around PhD theses supported by STFC funding or experiments on STFC large-scale research facilities. These records harvested from the British Library EThOS repository have been combined with records from STFC repositories, Researchfish (which is a UK Research Information Management platform) and from the research archive of the University of Oxford as an example of a university whose PhD students are often supported by STFC block student grants or by beamtime on STFC facilities. The bibliographic records from these repositories have been decomposed and topped up with records about organizations from GRID.AC[5] as a precursor for the forthcoming ROR[6] service. The resulted knowledge graph allows measuring the proliferation of different PID types in facilities research, also to explore opportunities for novel services using PIDs.

[5] https://www.grid.ac/
[6] https://ror.org/

3 Use cases

3.1 Use case: Persistent Identifiers for physical samples (IGSNs)

FREYA Partner: **PANGAEA**

Expanding the functionality of IGSNs for users of physical samples

PANGAEA is a data publisher for earth and environmental science located at Bremen University, which also hosts the core repository of the International Ocean Drilling Program (IODP). This core repository contains more than 150 km sediment core from the Atlantic Ocean, and more than 50000 samples are annually collected from the sediment cores for research purposes (Figure 1). The cores and samples are assigned a persistent identifier in the form of an IGSN, the International Geo Sample Number. The IGSN is an alphanumeric code that uniquely identifies samples from the natural environment and related sampling features.

Figure 1 IODP BCR sediment core repository in at Bremen University, Photo: © Marum

IGSNs are for the most part registered with additional metadata about the sample, such as date, sampling location and repository, which can be accessed by the user.

In a previous deliverable, D4.1, we demonstrated how PANGAEA has successfully integrated IGSNs in some of their published datasets, allowing the IGSN to be displayed in the metadata, along with other essential persistent identifiers, such as ORCIDs for authors, DOIs for articles and DOI for the dataset itself. Thereby, we enabled a PID Graph that facilitates access to essential metadata about the sample via the IGSN persistent identifiers and provides related linked resources via persistent identifiers in the dataset metadata description. However, it requires the user to manually “click” their way through the actionable PID-links available and compile the data and metadata they are looking for. Hence, the current set-up only provides a manual PID graph functionality and the information available takes its starting point at the datasets available through PANGAEA, which have IGSNs to identify samples. However, additional information available from sources outside PANGAEA are only accessible through implemented PID links, such as for example links to ORCIDs for authors and DOIs for publications.

The next step, which we describe henceforth, is to advance the PID Graph functionality to take a starting point at the physical samples and its IGSN number, and simplify the metadata collection for the user by (1) automating the metadata collection, (2) expanding the metadata search to also include external

resources and (3) helping the user to find resources from related IGSNs, such as for example sediment cores from the same area.

For this deliverable, we have built a use case around the Bremen Core Repository. Users of the Bremen Core Repository can get sediment core sections or samples from a core delivered for their research, either in a laboratory at Bremen University or in their own laboratory upon request. We want to enable the users working with cores from the repository to have easy access to data and metadata about the specific core that they are working on. This sediment core or sample has a barcode that contains related metadata including the IGSN. This is our entry point to the PID graph.

To enable this increased PID graph functionality, we will develop a software application; "an APP", that can browse the PID graph and collect information. In this deliverable, we design the FREYA IGSN Barcode scanner (FREYA IGSN Scanner) and describe its functionality. The realization of FREYA IGSN Scanner is planned as a task within PANGAEA for 2019/2020. For the users, we want to expand the PID graph functionality to encompass the following:

1. automated compilation of metadata for a specific IGSN;
2. expanded metadata collection to include sources outside PANGAEA;
3. facilitated access to related IGSNs.

FREYA IGSN Scanner will explore the following PID graph (Figure 2).

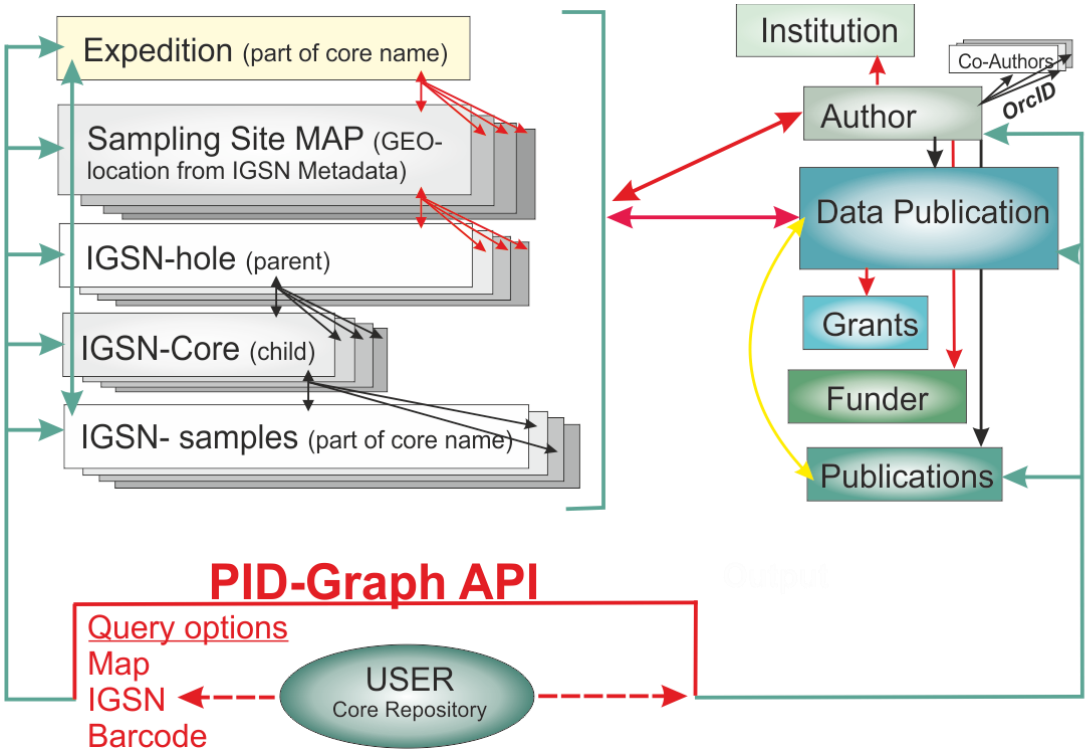

Figure 2 PID Graph development schema. The PID Graph will be navigated by a user along the green lines, indicating the possible entry points into the graph. Red connections are not in place yet; yellow connections are in place but cannot be navigated; black connections are in place and can be used to build the PID Graph.

Automated compilation

FREYA IGSN Scanner will include automated compilation, which will ease the users' access to metadata. Until now it has been up to the user to actively use the IGSN to find linked PIDs. With FREYA IGSN Scanner, we will automate this search, applying APIs to enter the PID Graph and automatically compile information starting with related datasets and articles from various sources.

Extended metadata collection

FREYA IGSN Scanner will initiate the search for related data and metadata using PANGAEA and the PID Graph incorporated in the PANGAEA database, which also links to expedition, funders and institutions. Since located at the same institution, there is a close link between the Bremen Core Repository and PANGAEA and both data and metadata from specific research cruises collecting cores for IODP are stored in the PANGAEA database, where it has its own portal, http://iodp.pangaea.de. For FREYA IGSN Scanner, this is a key entry point to the PID Graph. Furthermore, we will expand the PID Graph functionality to go beyond PANGAEA and also explore external PID resources. This will be demonstrated through a linkage to resources in FREYA partner DataCite.

Facilitate access to related IGSNs

The collection of sediment cores for the Bremen Core Repository has a hierarchical structure. One research cruise collects multiple sediment cores at different locations in the ocean. At each sampling location, there are multiple bore holes, from which the sediment cores are drilled out of the seafloor. As there is a limit to the length of each sediment core, each borehole can contain multiple sediment cores. After curation in the core repository, samples can be taken from the cores and sent out to scientists around the world. Individual unique IGSNs are assigned to bore holes, cores and samples. Furthermore, expeditions have their own identifier in the form of a number; although not a PID, the expedition numbers are also registered with the dataset metadata in PANGAEA.

For the user of the core repository, it is of great value to also get access to data and metadata from related cores, such as cores from the same or closely located bore holes or other cores from the same expedition. Furthermore, the user might also have interest in additional information such as data and articles from related cores and contacts to people working on related cores. FREYA IGSN Scanner will facilitate access to this information.

APP design

The APP has three levels (Figure 3): The IGSN entry, browsing the IGSN hierarchy and a results-level. The IGSN entry level allows the user to "enter" a relevant IGSN number. The IGSN hierarchy-level allows the user to browse through the IGSN hierarchy to find the relevant level with sufficient information and the results-section will show a compilation of related datasets and articles.

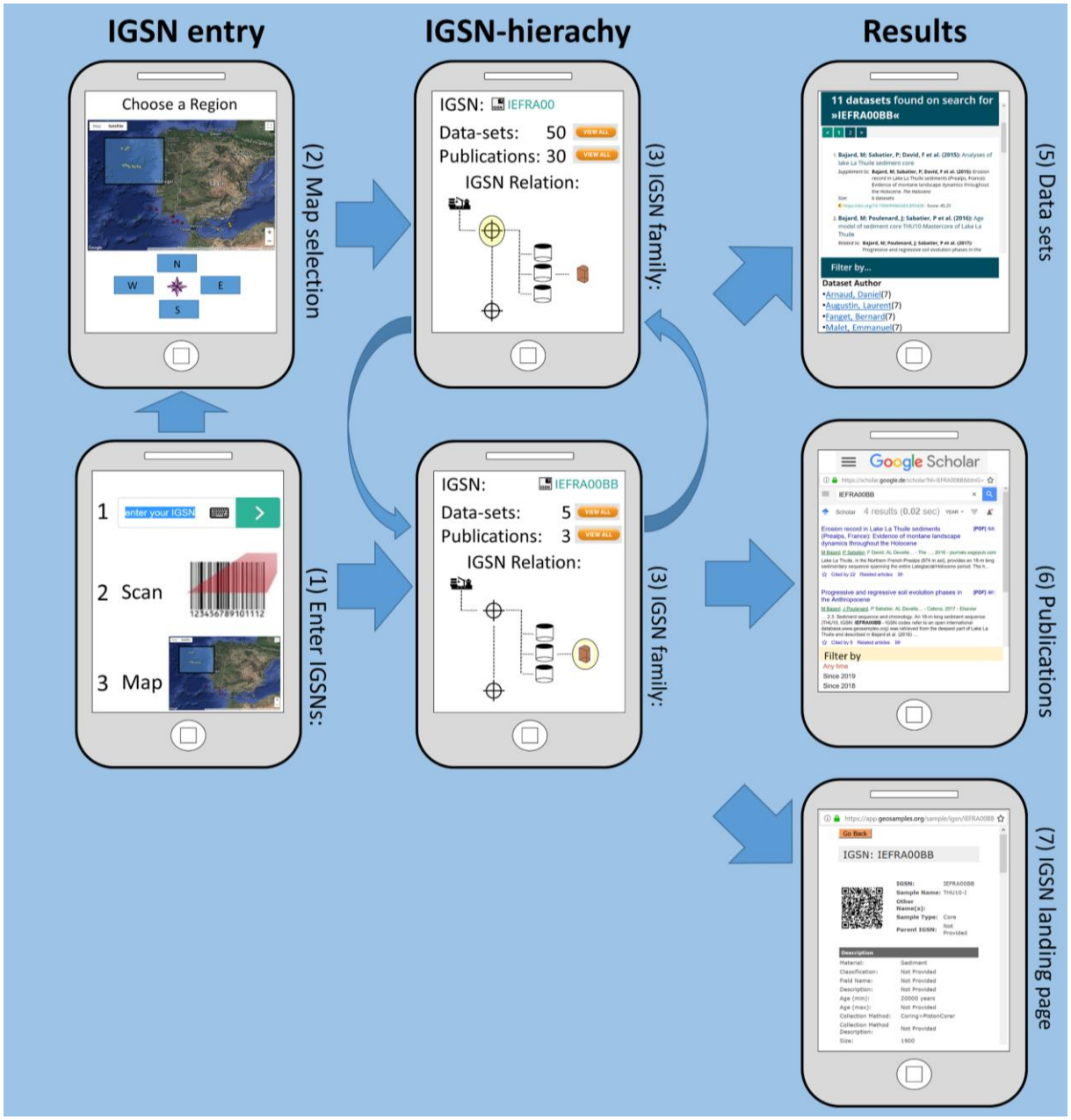

Figure 3 Workflow for the IGSN APP for Bremen Core Repository showing the pathway from IGSN submission to related Data collection, publications collection and corresponding IGSN landing page

IGSN entry sites

The entry site of FREYA IGSN Scanner (Figure 4 left) gives the user three options to search for information related to the IGSN. The IGSN-number can be entered directly or the barcode, located on the core or sample, can be scanned using a regular cell phone. As a third option, a geographical selection of IGSNs will be possible working from a map (Figure 4 right). It will be possible to select a specific region from a world-map, showing IODP-bore holes (sampling sites) from various expeditions.

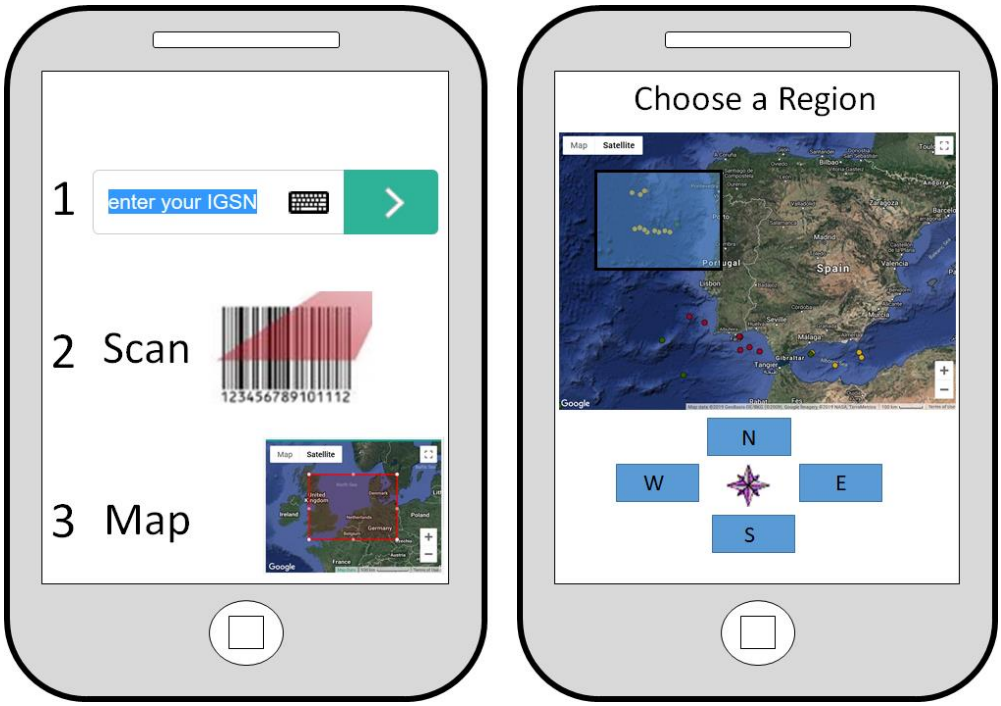

Figure 4 IGSN Entry sites (Left). Give IGSN by (1) type in the IGSN number (2) Scan a barcode on the cores (3) select IGSNs from map. Map-selection (Right), draw the area in question on the map.

The IGSN Hierarchy sites

In the IGSN hierarchy sites, it will be possible to browse through different levels of the IGSN hierarchy. The highest level is the expedition (), which can collect sediment from multiple bore holes (). There can be multiple sediment cores collected from each borehole and multiple samples can eventually be taken from individual sediment cores and used for research. Although "expedition" identifiers are not an official PID, we use the IODP expedition numbers, as it is part of the PANGAEA metadata.

The IGSN Hierarchy sites in the APP will allow the user to browse through these levels and see the amount of related datasets and publications at each level. As an example, if a user scans a sediment core from the Bremen Core Repository, they will be able to first browse to the IGSN for the borehole. From this level, they can display the related dataset and publications. Secondly, they can browse one level further back to the expedition. From this level, they can display the related dataset and publications and browse to other boreholes from the same expedition and get access information about geographically related cores.

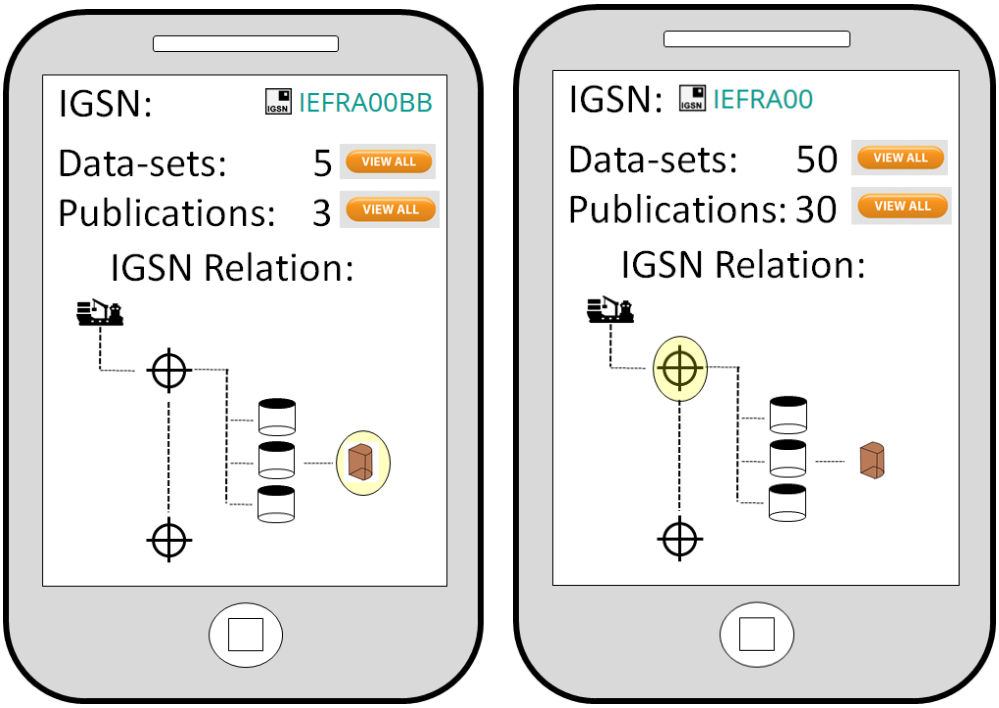

Figure 5 IGSN-hierarchical sites. The hierarchical relation between IGSNs for samples , cores , boreholes and expeditions is displayed. Each icon is "clickable" allowing the user to navigate through the related levels.

Result sites

The result site compiles related datasets (Figure 6 top-left) using the PANGAEA database. In this case, there will be a filter function allowing data sets to be filtered by author. In the cases, where the author has an ORCID, then it will be possible to resolve to the author's ORCID landing page.

The result site also compiles publications (Figure 6 top-right) using Google Scholar, which allows for article searches using IGSNs. Various filter functions are available through the Google Scholar search engine.

Additionally, the result section provides a link to the landing page of the IGSN, where there will be additional information such as material, age and collection method.

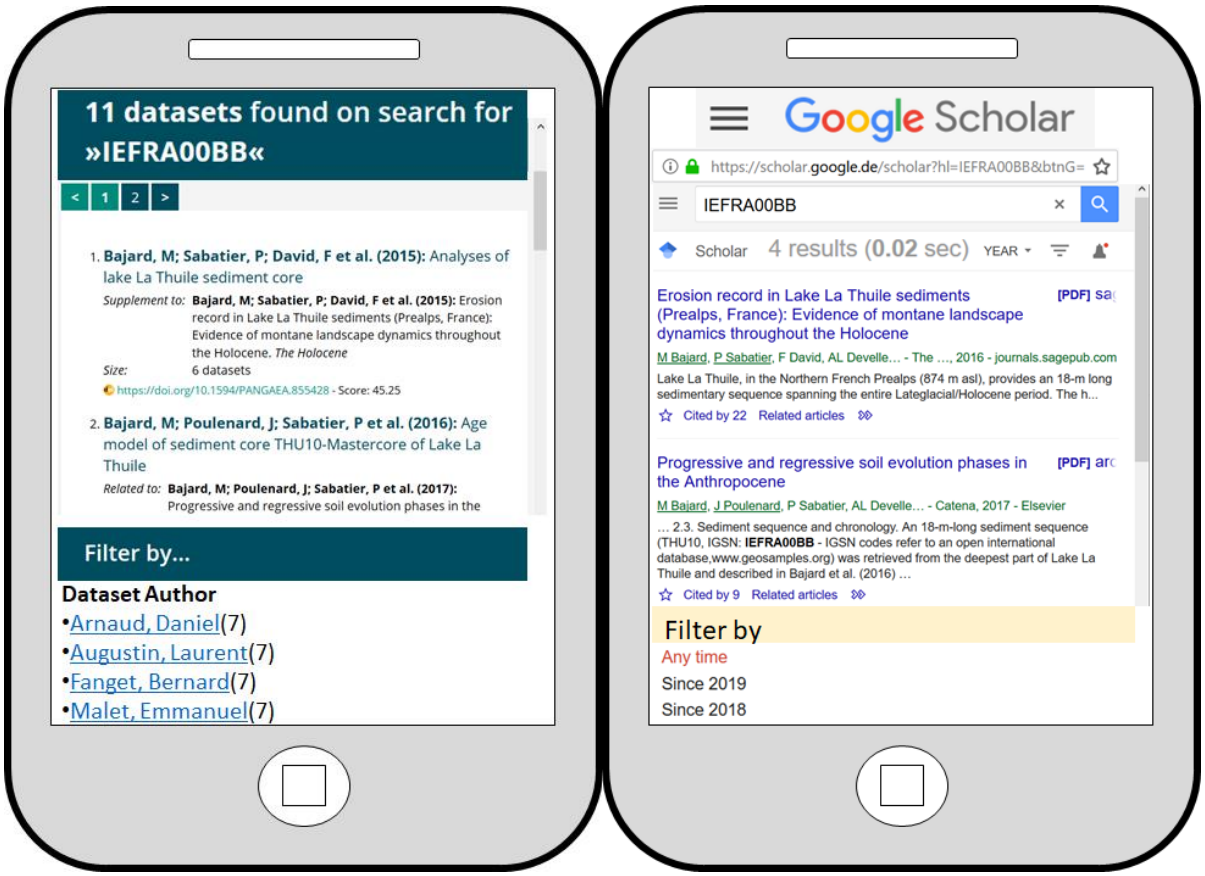

Figure 6 Result sites showing the data set (top-left), publications (top-right) and IGSN-landing page for a given IGSN (bottom)

Perspective

FREYA IGSN Scanner functions as a demonstrator, showing the benefits of linking IGSN-related information entities via their PIDs in a PID Graph environment and developing software that can help the user navigate the PID Graph. This first version of FREYA IGSN Scanner takes a starting point in the IGSNs assigned to sediment cores in the Bremen Core Repository. It accesses metadata stored on the IGSN landing page and in the PANGAEA database, as well as metadata stored externally at FREYA partner, DataCite, linking related authors and publications. It is a future goal to expand the reach of FREYA IGSN Scanner, building an API that can go beyond PANGAEA and FREYA partners and facilitate global searches for IGSN-related metadata and their associated research items. This would also include the implementation of additional PIDs, for example for research cruises and sampling locations. PANGAEA is currently in the process of finding programming capacity to realize this application. It is the goal to have FREYA IGSN Scanner working on a regular cell phone and in this way be a practical tool that can also be used while working on the cores in the laboratory. (After this deliverable was submitted, a prototype of FREYA IGSN Scanner app was built.

Update per November 2019

Past the submission of this deliverable, the first prototype of the FREYA IGSN Scanner was built. The APP works both for laptops and mobile phones. In its current version (1.0), the IGSN Scanner uses the SCANDIT barcode Scanner to identify the IGSN number embedded in the barcode, and the DataCite JSON REST API is applied to explore the PID graph around the identified IGSN number. This allows for identification of linked 1) Authors, 2) Articles, 3) Data-set, 4) Funders and 5) related IGSNs. The APP was demonstrated at the FREYA Midterm review in Luxembourg in July and is available here https://dataportals.pangaea.de/freya/igsn/)

Summary table

Pilot Application	FREYA IGSN barcode scanner (FREYA IGSN Scanner) - Enables users to scan a barcode on physical samples in the Bremen Core Repository to retrieve the IGSN number and search for associated resources, including related sample information.
Advanced PID graph functionality	The IGSN scanner allows users to search for resources associated with a specific IGSN using the PID graph, which include Scientists (OrcIDs), Funders (Crossref funder ID), data-sets (DOI), publications (DOI), and related samples (IGSN)
Method	• FREYA IGSN Scanner is a simple mobile web page • It uses the SCANDIT barcode Scanner and runs in any modern browser. • FREYA IGSN Scanner extracts IGSN IDs from the scanned barcode. It detects several formats like hdl.handle.net, igsn.org and other URL types. • Once an IGSN is found, FREYA IGSN Scanner starts a query to the DataCite JSON REST API, using the IGSN as lookup key for dataset relations. • It formats all related datasets as a citation list, using the DataCite citation formatter. • It extracts all persons/scientists with ORCIDS, other IGSN identifiers and funding references. • It extracts DOIs of publications • If available it identifies funders through dataset publications via their Crossref Funder ID

Rationale	At the Bremen Core repository more than 150 km sediment core from the Atlantic Ocean is stored and available to scientists for use in their research, and 50.000 samples are handed over to scientists each year. Each core has a barcode referring to its IGSN number. Through the development of the FREYA IGSN Scanner-app, we provide the user with easy access to all the information that is linked to a the IGSN-number. This improves the findability and accessibility of the available information as well as facilitates reuse of existing data in accordance with the FAIR principles. PIDs are envisaged to play a primary role in the EOSC in particular in regards to the PID Handler, which is a class of services with in EOSC to generate and resolve Persistent Identifiers (PIDs). The FREYA IGSN Scanner app was specifically developed as a service to accommodate functionality envisioned for the EOSC PID- Handler, namely, the resolution of PIDs. Additionally, the FREYA IGSN Scanner app takes the functionality one step further enabling both the discovery and resolution of related PIDs.
Target USER	Researchers working with physical samples, which have IGSNs assigned for identification. This would include, for example, researchers working with sediment samples from the Bremen Core Repository.
EOSC	FREYA IGSN Scanner contributes to the EOSC specified services for researcher in particular in regards to the “search and browse”, “recommender” and “repository service”, leading researchers to specific repositories with related data or articles and recommending related IGSNs. Furthermore, building the PID graphs, this pilot application also contributes to the “PID handler”. However, this pilot application is currently at TRL 6 and does not yet qualify for the EOSC service catalogue. The catalogue is restricted to services with TRL 8 and 9, including only services considered mature enough to be at production level. Hence, implementation of FREYA IGSN Scanner in the EOSC service catalogue is not yet realizable.

3.2 Use case: Persistent Identifiers for Documents (DOIs)

FREYA Partner: **British Library (BL)**

Using DOIs to streamline access to resources across multiple locations improving discovery and access to the British Library's collections

The British Library's catalogue holds approximately 82 million records for books, journal articles, sound recordings, newspapers, printed maps and musical scores, however a proportion of these records are in fact duplicates of the same resource but which have been acquired through a different method. The British Library will improve the discovery and access to resources, such as journal articles, held across multiple locations in one place through the use of an API, which matches DOIs to multiple instances of the same resource and present all the potential ways to access the resource in one easy to access interface. This will make it easier for users to find the right resource for them depending on where they are and how they want to access the material.

The new service, BLDOI, will improve discovery of the BL's resources, primarily journal articles, by ensuring all access options for an article are displayed within one record. This work supports the PID Graph by creating a one-to-many relationship between an identifier and multiple instances of the same resource. The methodology behind these developments comes from consistent user reports of difficulties finding resources held by the BL. This project was conceived as a simple method to improve the end user's discovery experience in an "easy-to-deliver" way.

Current Discovery Experience

At present a search in the BL's discovery service, Explore[7], an instance of Primo, a discovery tool used by many libraries, for a resource in the BL's catalogue can produce several identical results which contain different access options which is confusing to users, such as in Figure 7.

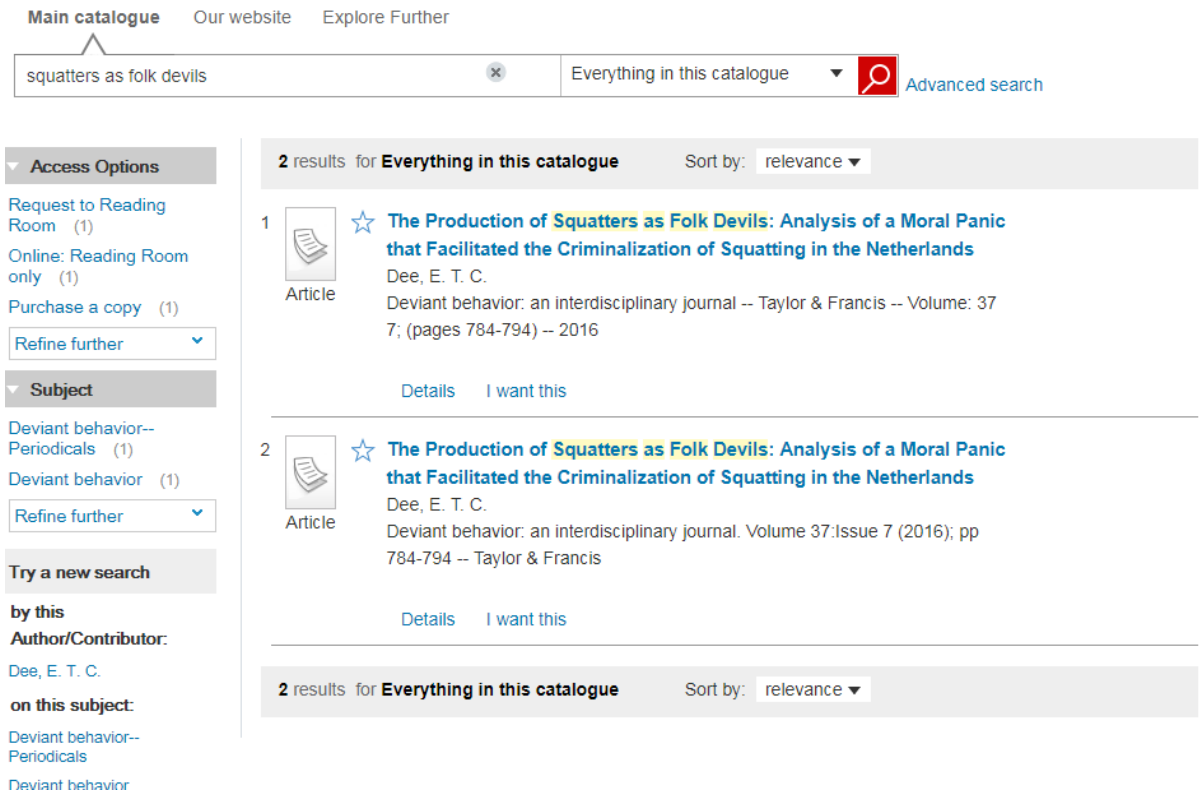

Figure 7 The results from a search for a journal article yielding two identical results.

When a user tries to access the article using the "I want this" link for each of the results, the options presented are different, see Figure 8 and Figure 9. In Figure 8 the access option displayed is to use the British Library On Demand Service, this is because the record is a purchased resource and only has information about the British Library On Demand service.[8] In Figure 9 only the access options for Legal Deposit[9] content are displayed because it uses the Legal Deposit Archive Resource Key (ARK) to display this access route.

[7] http://explore.bl.uk

[8] The BL's On Demand Service provides copies of books and articles for a fee. https://www.bl.uk/on-demand

[9] All publishers in the UK and Ireland are required to provide a copy of every publication to the British Library but this content can only be accessed within the Library's buildings. https://www.bl.uk/legal-deposit. The legislation behind this requirement stipulates that this collection cannot be used for profit, hence the need for an additional collection for the commercial On Demand service.

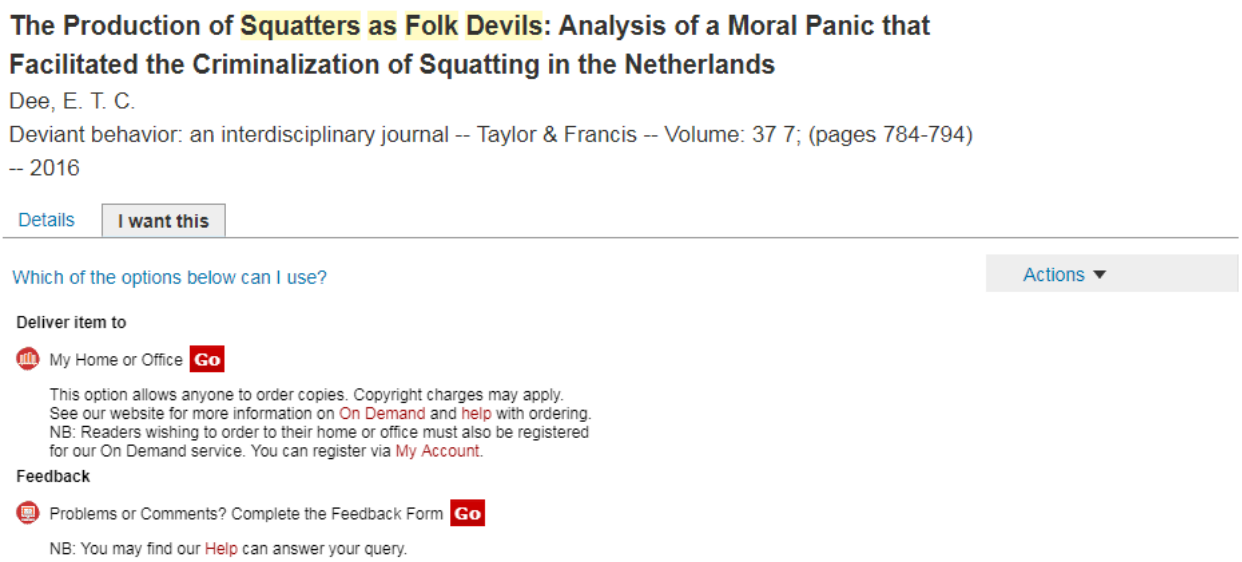

Figure 8 This result only displays the option to access the article using The British Library On Demand Service, which allows users to order copies of articles for a fee.

Figure 9 This result only displays access options for Legal Deposit content as that is the only information contained in this record.

This search can be expanded using the Explore Further option which includes metadata from an externally held subscription service, Primo Central Index, for resources held elsewhere. Additional records for the same article may be found in these results but as the Primo Central records contain none of the BL's identifiers, no BL access options are listed. For records with a DOI, SFX is able to list Open Access versions available using a service provided by Unpaywall[10].

Each item found links only to a specific digital item, making use of a system-specific identifier in the metadata record, Figure 10 provides a diagram of the options and the resultant text presented to a user. For each of the different routes the access options are different but a user needs to look at each record

[10] https://unpaywall.org/

to see these different options. In practice they do not look at all the options and it is particularly frustrating for users where content is available outside of the library building but the user does not find that record.

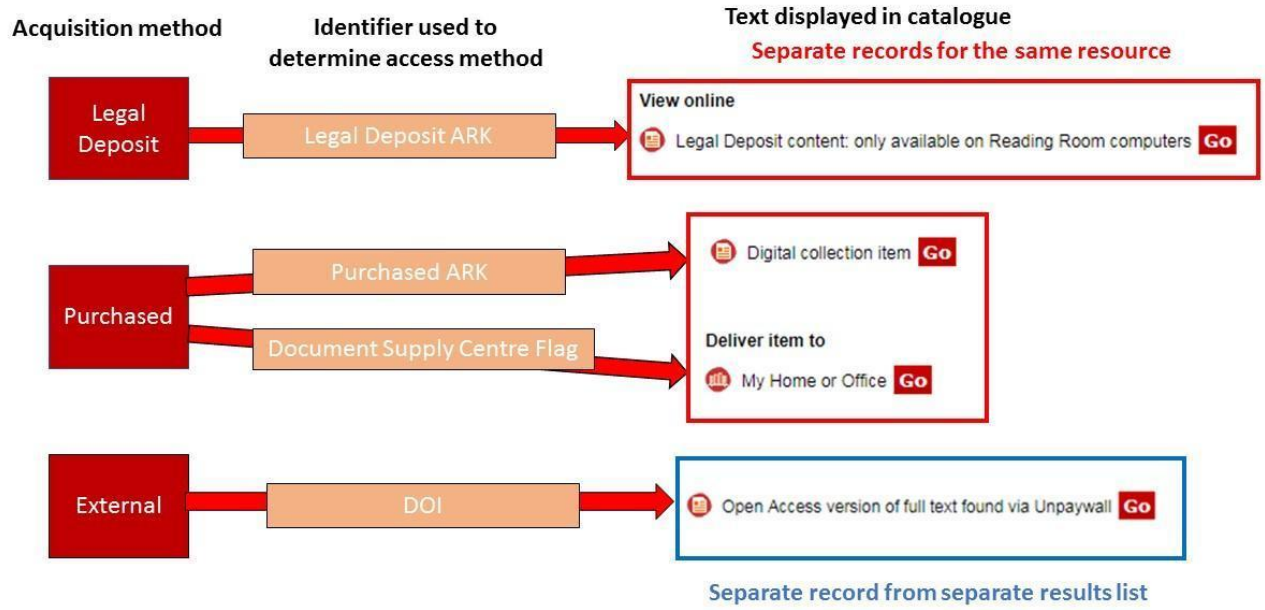

Figure 10 Diagram representing the current discovery experience. The different acquisition routes (left) provide different access text (right), using different identifiers, depending on the acquisition route. The different acquisition methods also lead to the creation of separate records for the same resource.

Desired Discovery Experience

Regardless of acquisition route, all records of the same resource have the same DOI and the planned development will use the DOI to ensure all access options are available to a user looking at one record, see Figure 11 for a diagrammatic explanation of what the service will deliver and Figure 12 for a mock-up of how the page will look. This means that items found in external subscribed indexes, such as Primo Central index will also successfully link to British Library services in addition to Open Access and other services.

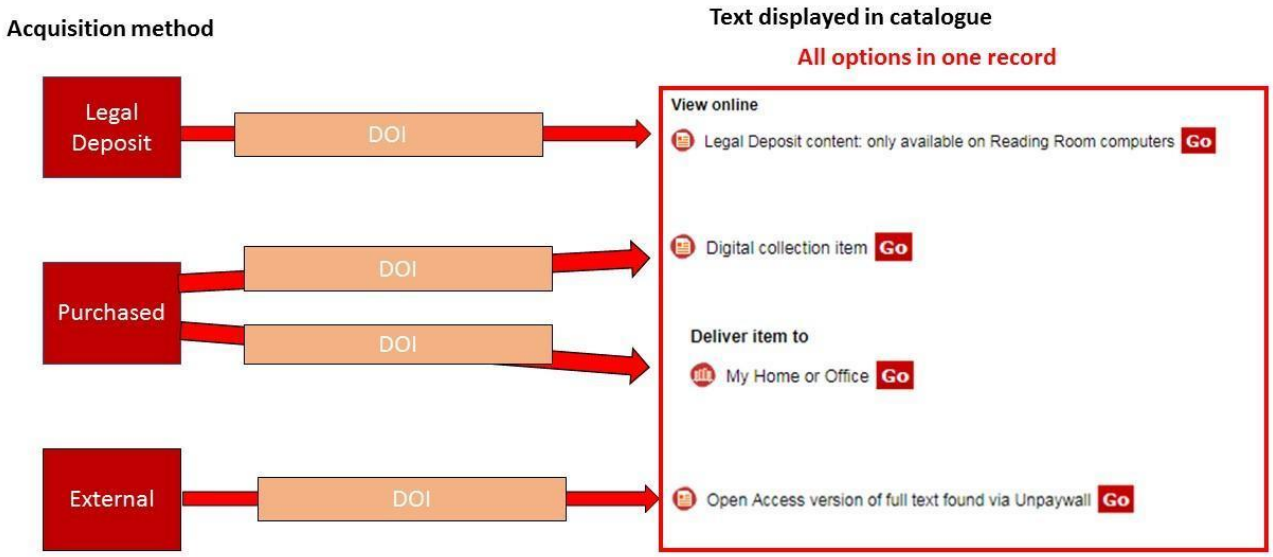

Figure 11 Diagram representing the planned development. All access options will display in one record allowing the user to make an informed choice of access route. The DOI will be used to match the results from the different acquisition routes to one record.

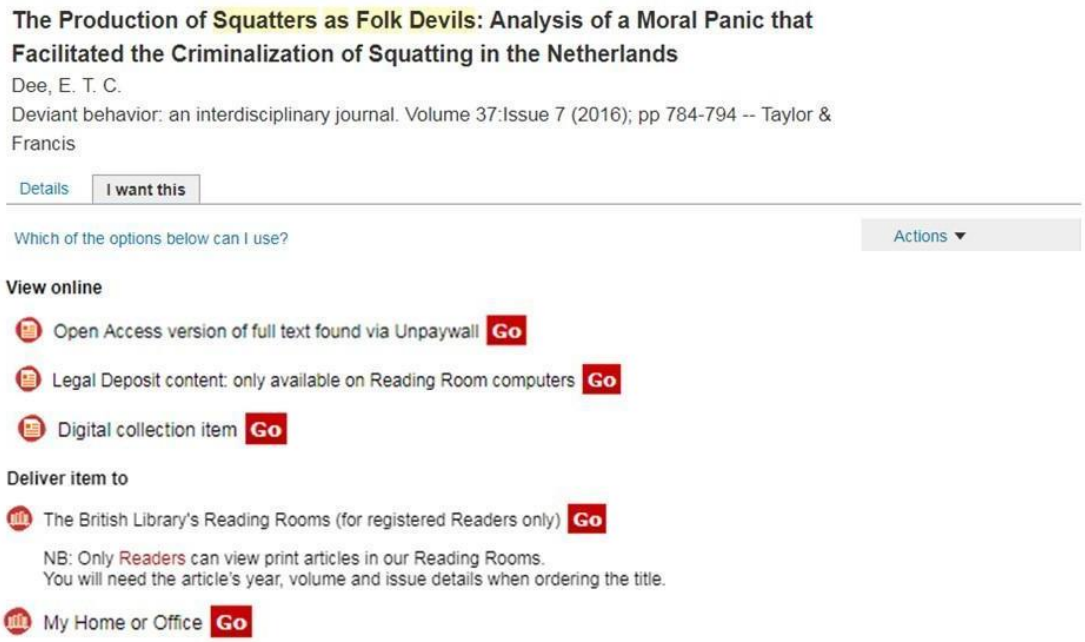

Figure 12 A mockup of how a record which provides all access options could look.

Future work related to this is planned to define and adjust the ranking of the access options depending on users' location or access status, i.e. in the Library's buildings or elsewhere, logged in or out. For example, if a user is not in the Library and not logged in, they will be presented with Open Access option (if available) as the top result. This will mean that users will receive a tailored discovery experience, making it quicker and easier to navigate to resources.

Once the developments outlined below have been implemented, the number of records held in the BL's index could be reduced because separate records for the same articles acquired through different means will not be required. This will allow the BL to focus instead on the unique content it holds and improving the discovery experience for users. The API created may have applications to other institutions, either those who harvest from BL collections or those who wish to rationalize their access to content. The BLDOI service provides a mechanism to streamline the data integration between the BL and other UK Legal Deposit Libraries.[11] It may also be a valuable tool for British Library On Demand institutional customers.

Planned development

To make the changes described above, a database will need to be created and populated to map DOIs against the related BL access identifiers (ARKs) to run a microservice, which will return ARKs relating to the item when queried with a DOI, creating a type of PID Graph connecting multiple items, which have the same DOI. As it is a simple lookup, it will be possible to build the microservice, so that it performs well under a high load. The API will resolve to access URLs for SFX to create menu links to display on the results page. An SFX target will also need to be built to consume the microservice. Figure 13 illustrates the process from a user's search to the results being displayed.

[11] https://www.bl.uk/legal-deposit/about-legal-deposit

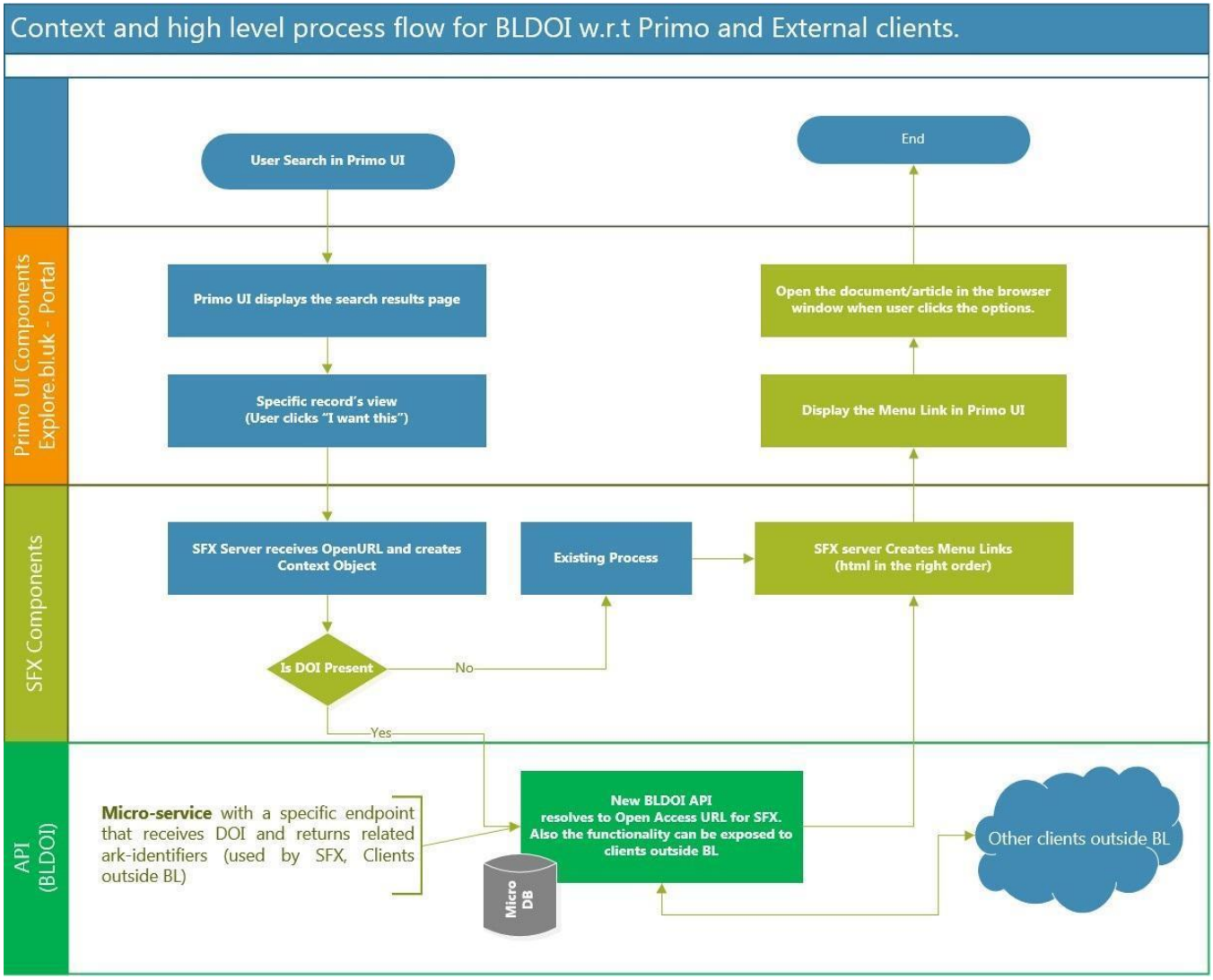

Figure 13 A swim lane diagram demonstrating what the planned developments (bottom lane) will deliver from the point of view of the process from a user's search to the link to the resource being returned to the user to click on. It illustrates how this new service (bottom lane) will interact with the existing process to display results. Other clients include other UK Legal Deposit Libraries and British Library On Demand institutional customers.

The required functionality of the API is, that it should accept a single DOI and look up the local database and return the relevant ARK identifiers, returning a not found message, if the DOI cannot be found. This should also be applicable for a DOI array or collection of DOIs. The API will accept requests and output in JSON format.

The ranking of access options will also require development work and this will be scoped in the coming months.

Development work on the BLDOI service is due to commence shortly, and it is due to be delivered before the end of 2019. The ranking of access options will be delivered in 2020.

Summary table

Pilot Application	Using DOIs to streamline access to resources across multiple locations improving discovery and access to the British Library's collections through the library's discovery tool Explore.
Advanced PID graph functionality	*Enabling access to content in multiple locations* The API enables access to all instances of a particular resource in one record on the library's catalogue allowing easier access. This allows users to access to the various versions of a resource in a simple and accessible way.
Method	• A database will be created that matches the DOI against the various ARK identifiers (used by the library), which are present for resources procured via different routes. The database will allow a one to many relationship between the DOI and ARKs to be created. • When a user searches for a resource, which has a DOI, an API, which will be created, will query the database and return all the ARKs associated with a given DOI. • The user is presented with a list of access options, which are ranked according to which type of user they are (registered onsite, registered offsite or unregistered), presenting the most relevant option first.
Rationale	The current configuration of the Explore tool means that users have to navigate through multiple duplicate resource records to find the version of a resource they can access. This new API will mean all access options will be presented in one place. While the exact configuration of the British Library API may not be applicable universally, the rationale could be taken up by other European research libraries.
Target USER	Researchers using the British Library's catalogue.
EOSC	BLDOI contributes to the EOSC specified services for researchers in relation to search and browse and recommender services. The API will be available for other libraries to use, however, it is as yet unclear how applicable this will be to other institutions.

3.3 Use case: CERN Analysis Preservation (CAP)—a community PID Graph

FREYA Partner: **CERN**

Improving automatic population of CERN Analysis Preservation records with information from external sources and development of visualization tool.

CAP Graph

CERN's pilot applications build on the presented work in D4.1 which highlighted several services which are part of the CERN environment. In this deliverable, we focus on enhancements for CERN Analysis Preservation (CAP)[12], a closed system and community PID Graph to preserve and identify research objects from the physics research lifecycle (physics analyses), e.g. data, code, workflows, contextual information.

A basic workflow had been established to enable the user to connect the access-restricted materials from the community PID Graph (i.e. CAP) to the public PID Graph. However, CERN Analysis Preservation aims at being a complete preservation service, capturing a complete image of a physics analysis. Hence, we considered that it would be beneficial to enable a (meta)data exchange from CAP to a range of publishing services to ensure that CAP is always up to date and preserves complete analyses and their context (including public metadata). On the same note, the established integration with ORCID was not yet very user-friendly and did not benefit from fetching more of the information available in the ORCID records.

The suggested advancements in CAP can benefit from better APIs and direct data access. To better illustrate the benefits to the user community, it was decided to add visualization tools showing direct data access, which is done via the integration of Python notebooks.

In summary, in this use case we will expand on the above concepts and particularly work on the following:

- Improve automatic population of CAP records with metadata from external sources, so that CAP records become more complete (archive) easily (without having to navigate to the external service to transfer information). This will be demonstrated by fetching information from ORCID and Zenodo as a first implementation.
- Develop a visualization tool for Python Notebooks illustrating direct data access in the PID Graph.

Regarding the fetching functionality, work has already been done to enable ORCID and Zenodo integrations and the aforementioned enhancements will build on top of that. Deliverable 4.1 provides more details on these implementations. More specifically, already available in CAP are:

- ORCID: Automatically finding and adding an ORCID ID to a CAP record when inputting a name of an author, provided one was already registered in ORCID for the name being entered.
- Zenodo: Fetching code from GitLab/GitHub and uploading code from the CAP User Interface (UI) to Zenodo to be published.

Adding this functionality of fetching and collecting (meta)data from various PID services and attaching them to our own records helps us advance our indexing and search infrastructure (e.g. searching all physics analyses that used a specific software with a specific DOI). Within the context of the PID Graph,

[12] https://github.com/cernanalysispreservation/analysispreservation.cern.ch

such features also enhance findability and interoperability of information, while also enriching the CAP PID Graph.

This functionality allows the users to provide a specific PID (or a service internal ID) through our UI component and attach the resolved (meta)data to a CAP record through using different APIs from a variety of services. By integrating this to work with our search engine (Elasticsearch) and creating specific mapping for these resources, we are able to provide an advanced browsing experience to our users.

The current implementation supports information fetching from ORCID and Zenodo. In the following, the workflow is described using screenshots for each step (Figure 14, Figure 15, Figure 16, Figure 17, Figure 18).

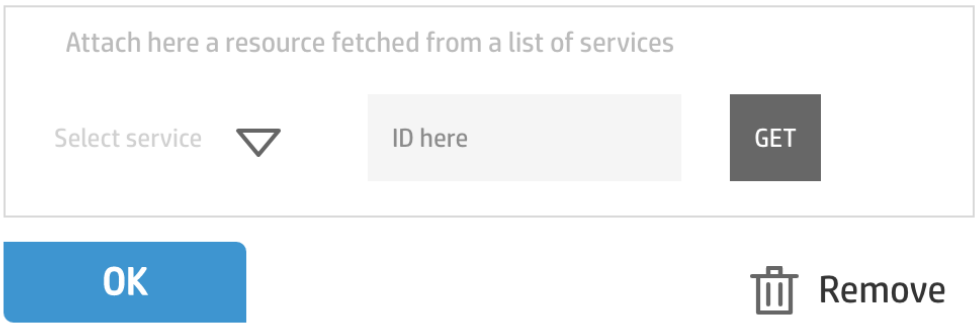

Figure 14 Fetching functionality in CERN Analysis Preservation (1) - UI component implementation of the functionality.

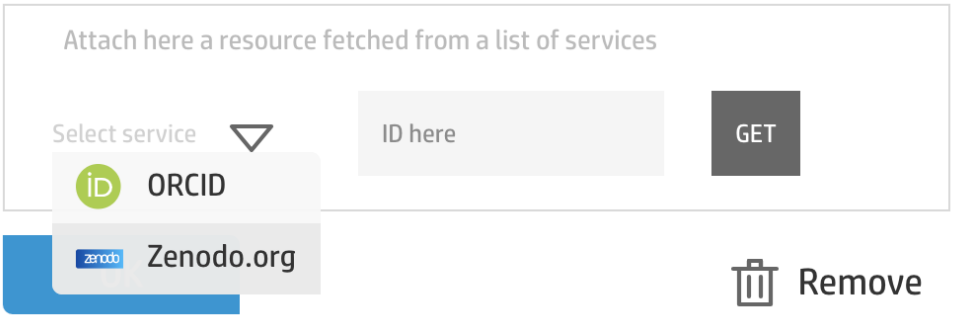

Figure 15 Fetching functionality in CERN Analysis Preservation (2) - The user can select a service to fetch from; once the specified ID/PID is entered, the information is attached to the CERN Analysis Preservation record.

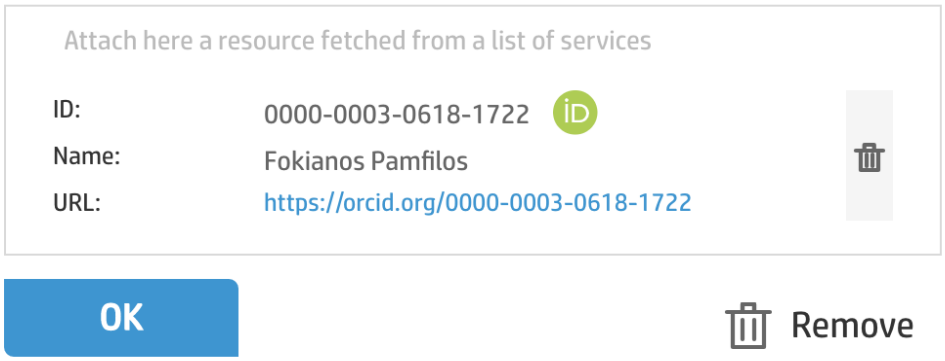

Figure 16 Fetching functionality in CERN Analysis Preservation (3) - Example fetching of an ORCID record (with a possibility to explore and expose more information and links from ORCID in the future).

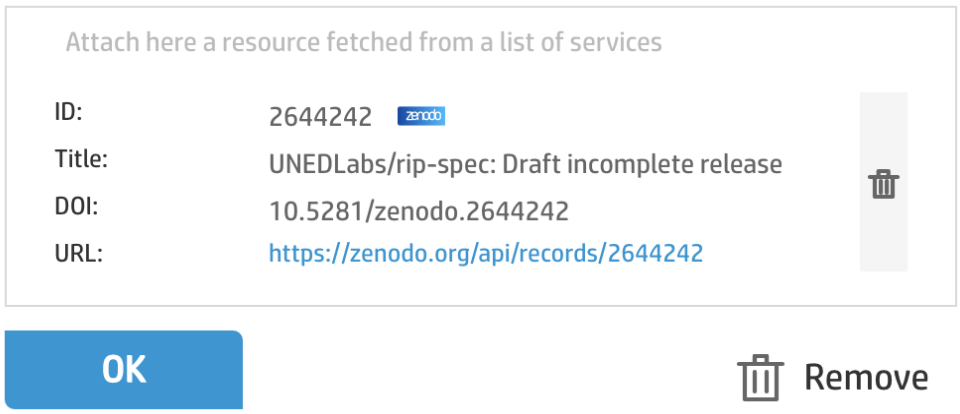

Figure 17 Fetching functionality in CERN Analysis Preservation (4) - Example fetching of a Zenodo record.

```
3 : {...} 2 items
4 : { 2 items
  "source" : { 2 items
    "service" : string "zenodo"
    "externalID" : string "2644242"
  }
  "fetched" : { 12 items
    "conceptdoi" : string "10.5281/zenodo.2644241"
    "conceptrecid" : string "2644241"
    "created" : string "2019-04-17T19:32:16.717714+00:00"
    "doi" : string "10.5281/zenodo.2644242"
    "files" : [...] 1 item
    "id" : int 2644242
    "links" : {...} 9 items
    "metadata" : { 12 items
      "access_right" : string "open"
      "access_right_category" : string "success"
      "creators" : [ 1 item
        0 : { 2 items
          "affiliation" : string "UNED"
          "name" : string "Luis de la Torre"
        }
      ]
      "description" :
      string "Specification of the Remote Interoperability Protocol"
      "doi" : string "10.5281/zenodo.2644242"
      "license" : { 1 item
        "id" : string "other-open"
      }
      "publication_date" : string "2019-04-17"
      "related_identifiers" : [ 2 items
        0 : { 3 items
          "identifier" :
          string "https://github.com/UNEDLabs/rip-spec/tree/0.35"
          "relation" : string "isSupplementTo"
          "scheme" : string "url"
        }
```


Figure 18 Fetching functionality in CERN Analysis Preservation (5) - Part of the JSON from a CERN Analysis Preservation record. Displayed is archived and searchable data fetched from Zenodo.

This work on enhancing the fetching feature in CAP will enhance the PID Graph functionality by making the user experience more seamless as content from multiple sources will be directly accessible through a single entry point (the CAP record). The user will be able to fetch information and populate their analysis record without having to do it manually, which will make it that much easier for them and therefore more possible to include such metadata.

Visualization tool to illustrate direct data access and the CAP graph to users

CERN Analysis Preservation is an internal community PID Graph by itself. However, as users are encouraged to submit and connect various research materials, their potential and connections are not obvious to them; building a visualization tool addresses exactly that.

The following implementation integrates Python Notebooks directly in the user interface of CAP. The Python Notebooks are an interactive computational environment, in which you can combine code execution, rich text, mathematics, plots and rich media. It is then possible to directly exploit the direct access via our REST API. The notebooks fetch record information in JSON format to then use it in the containerized image. This approach helps creating a parameterized visualization to explore the analysis output live. This way, the users are able to see and explore the results of a HEP (High Energy Physics) analysis is through the "likelihoods", as they are shown in Figure 19 and Figure 20.

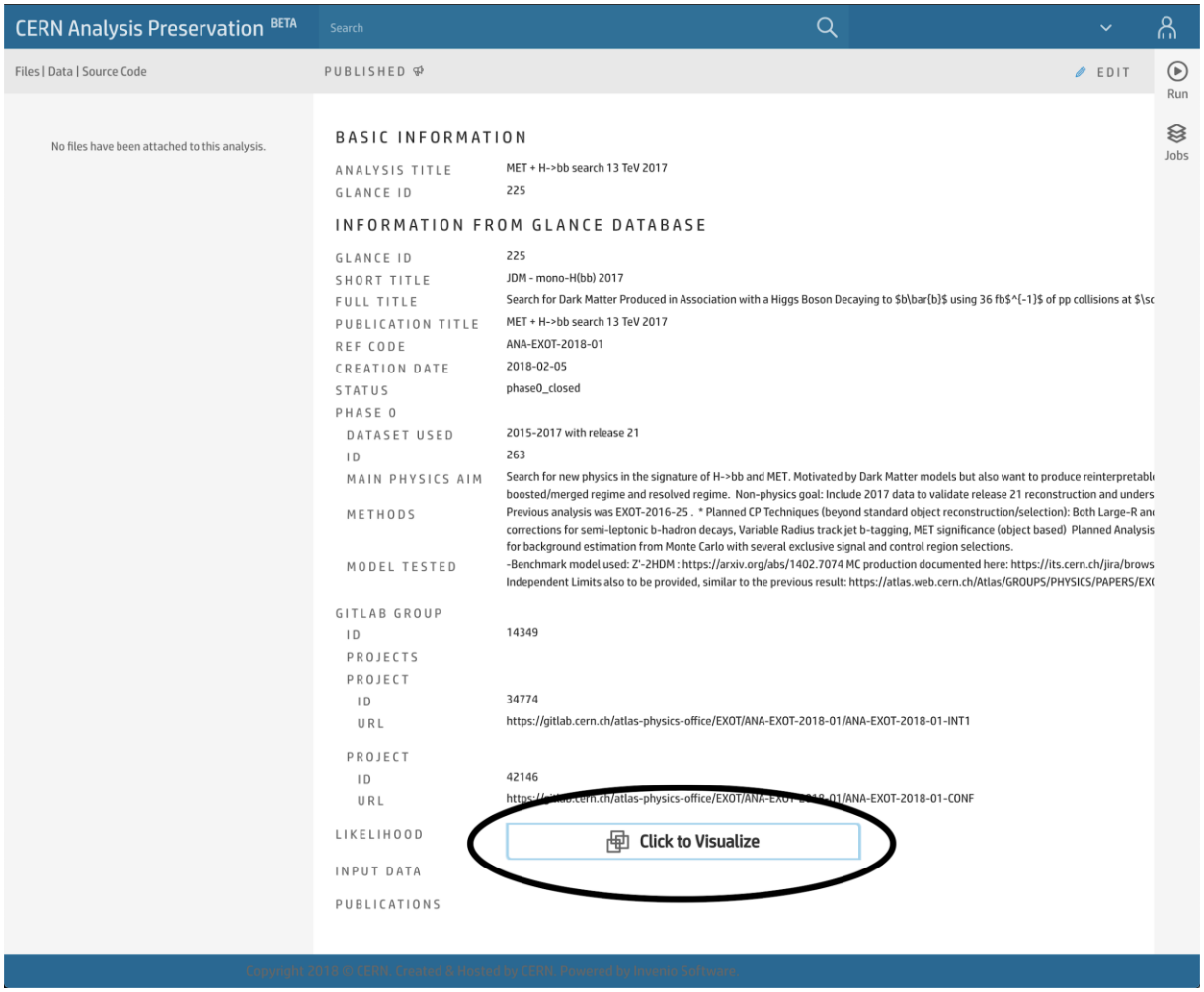

Figure 19 A CERN Analysis Preservation record that provides the functionality for visualization (see circled button at the bottom of the page).

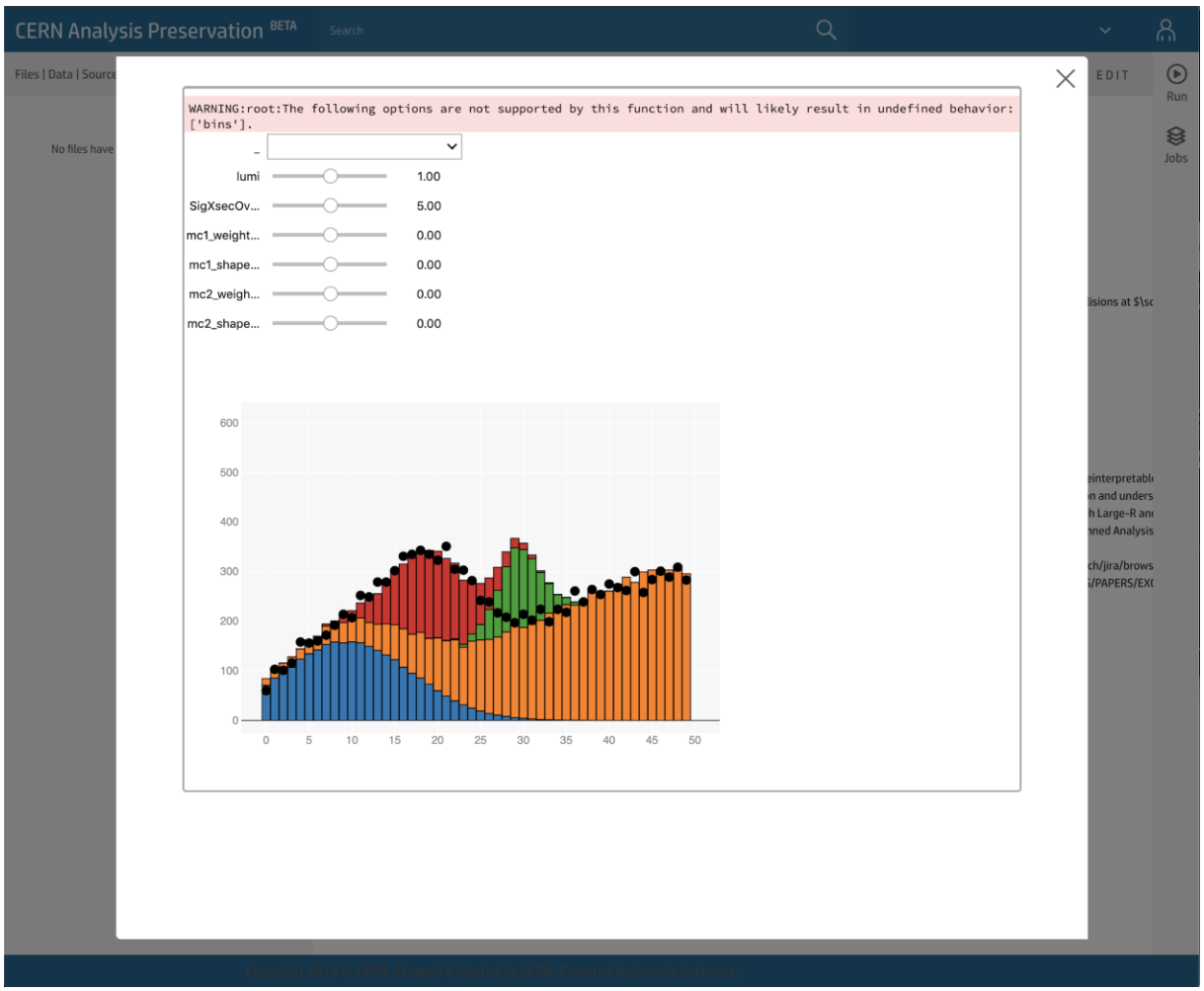

Figure 20 The user interface on CERN Analysis Preservation that enables users to use a Python Notebook to explore the data in the CAP PID Graph. This happens by fetching data from CERN Analysis Preservation's REST API. The example above visualizes likelihoods from an example analysis.

By pressing the button (Figure 19), the Python Notebook then uses the direct access via the REST APIs to fetch data from the CAP graph. The results from one example analysis are given above (Figure 20).

The CAP visualization features are already implemented in CERN Analysis Preservation and they will be available in the production system in the coming weeks. In the future, the CAP graph could be expanded so that CAP is able to fetch information from other CERN services, such as INSPIRE, CERN Open Data, CDS, as well as non-CERN services.

These integrations could also be used as an example for other services outside the High-Energy Physics environment, since the actual implementation is discipline-agnostic.

Summary table

Pilot Application	Improving automatic population of CERN Analysis Preservation records with information from external sources and development of a visualization tool.
Advanced PID graph functionality	The fetching and visualization functionalities are meant to enable the user to benefit more from the service, through enriching the CAP PID Graph with more information/resources and easier ways to access it.

Method	• Improve automatic population of CAP records with metadata from external sources, so that CAP records become more complete (archive) easily (without having to navigate to the external service to transfer information). This is demonstrated by fetching information from ORCID and Zenodo as a first implementation. • Develop a visualization tool for Python Notebooks illustrating direct data access in the PID Graph.
Rationale	The development of CERN Analysis Preservation addressed the need of having a dedicated tool at CERN for preserving analyses in a sustainable way that adheres to current Open Science standards as much as possible (focusing on the 'R' in FAIR). CAP is a closed-access, because active analyses cannot be public and may only be accessed by a specific person/team/collaboration. However, all the source code is available on GitHub, and various PID integrations are already in place (e.g. fetching information using DOIs or ORCID iDs or pushing code to Zenodo from CAP to get a DOI). CAP can connect information from various CERN and external information sources to make analysis records as complete as possible (creating its own PID graph). It ensures that analyses are easily reproducible through the connection established to the REANA service (reproducible research data analysis platform). Currently, CAP is tailored to the HEP community, but as it grows it has the potential of being used by researchers outside of CERN and possibly in other disciplines.
Target USER	Primarily CERN physicists. In the future, this tool could be also used by other disciplines for preserving analyses (with some modifications), as the technology used is discipline-agnostic.
EOSC	CAP is still in a pilot stage and cannot be included in the EOSC service catalogue at this time. Also, in order for the tool to be useful for the users at CERN that want to preserve their analyses, it needs to be restricted-access. In general, tools like CAP could contribute to various EOSC services for researchers, possibly "Repository", "Peer Review", "Search & Browse", "Workflow", "Workspace", and more.

3.4 Use case: Exposing provenance information around preprints

FREYA Partner: **EMBL-EBI (Europe PMC)**

Extending access to the literature PID Graph by exposing provenance information for preprints

Versioning functionality associated with indexing preprints in Europe PMC

Article versions have become more common with the introduction of preprints and open peer-review publishing models such as those powered by Open Research Central.

In the life sciences, peer reviewed journal articles are the means by which research results are shared, by which research impact is assessed and used to drive the procurement of subsequent funding and career progression. In the past few years the use of preprints has become increasingly popular. These are non-peer reviewed articles, posted to preprint servers and generally viewed as precursors to published articles. The reasons for making preprint versions available ahead of peer review are many, and include vastly reducing the time taken from having research results available to public sharing (instead of enduring the embargo imposed by formal peer review, publication and time to sharing beyond journal subscription holders) ; as well as being able to claim early primacy for exciting research results.

Europe PMC began indexing preprints in July 2018 and in so doing broadening the scope of literary content in its repository. Importantly, where indexed preprints have a matching peer-reviewed journal article, these are linked within Europe PMC (see Figure 21). This functionality begins by addressing the user story: “Linking preprints to their published article version: As a data scientist (or researcher), I want to know whether any given preprint has subsequently been published. If yes, then for these to be linked reciprocally (from preprint to publication; from publication to preprint)”. Discoverability of the preprints is enhanced further by being able to link indexed preprints to one’s ORCID profile.

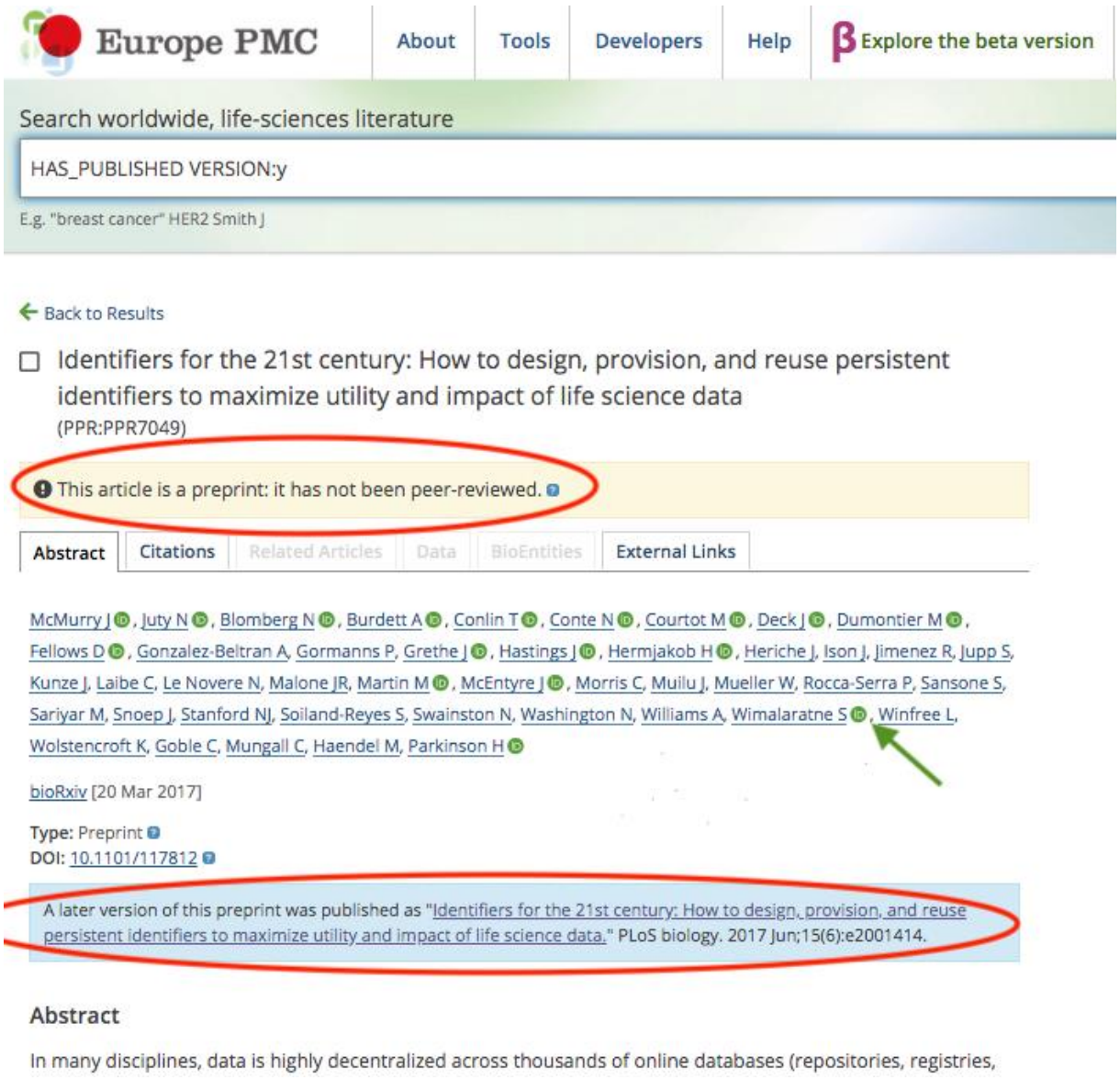

Figure 21 A preprint indexed by Europe PMC as it appears in the user interface: It is clearly designated as preprint (yellow bar) that links to a peer-reviewed version published in a journal (blue bar) and to 17 ORCID profiles (ORCID icon indicated by green arrow). The search syntax: HAS_PREPRINT:y was used here, and limits search results to peer reviewed articles which were preceded by a preprint version of the article. The search syntax, HAS_PUBLISHED_VERSION:y, limits search results to preprints which were later peer reviewed, edited, typeset and published as journal articles.

As a result of linking DOIs to ORCID iDs, author profiles generated by Europe PMC display content that has been claimed to a specific ORCID iD including preprints - see Figure 22.

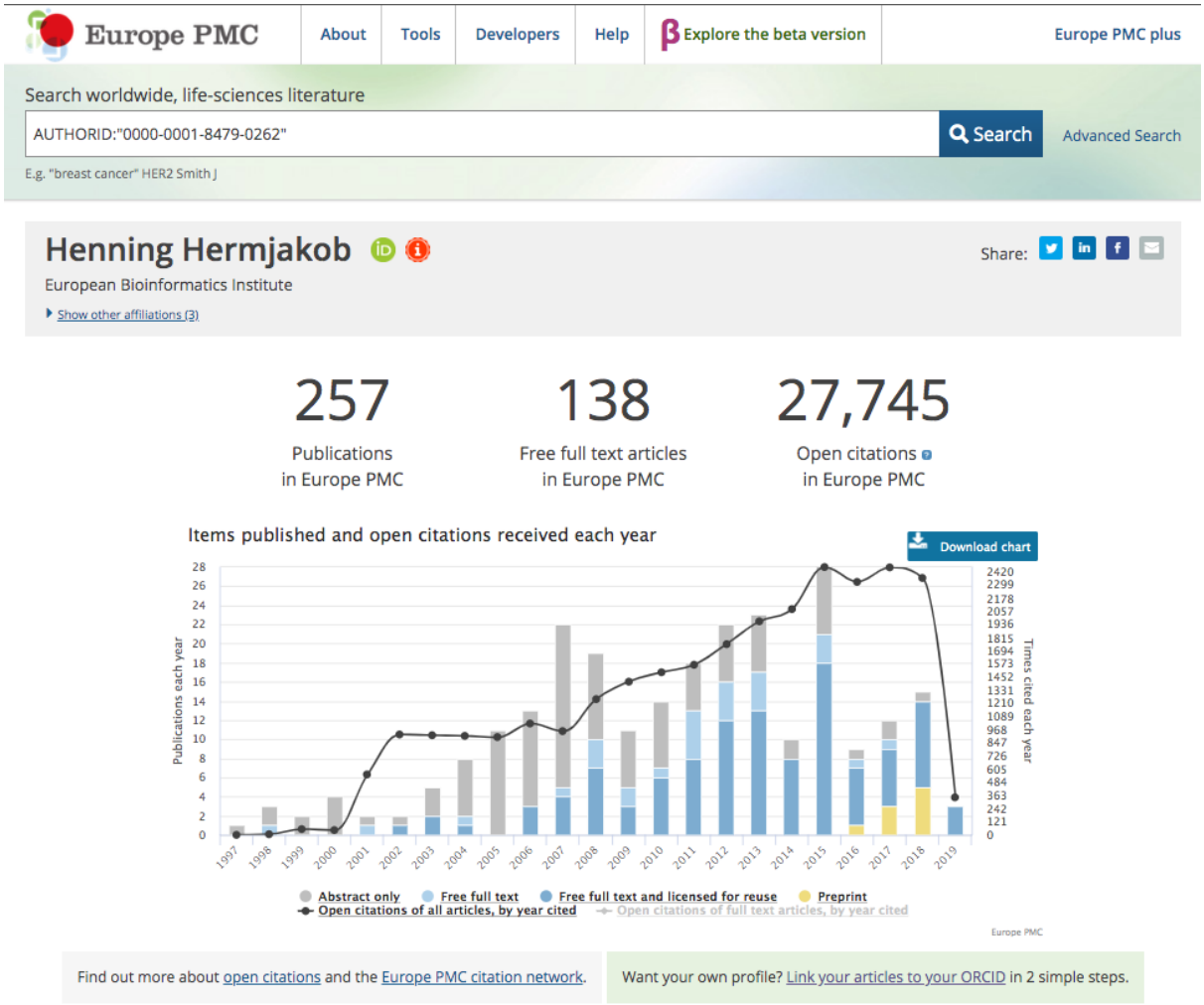

Figure 22 An author profile chart generated by Europe PMC. The profile generator displays the publications that an author has linked to her/his ORCID iD.

In this case, the author has 258 publications, for which 27745 citations by other articles can be counted; the total number of citations obtained each year is indicated by black dots. (To be counted as an open citation, the citing articles must be uniquely identified e.g. by resolving to a PubMed ID). Of the total number of publication records listed, a total of 138 are available as freely accessible full text articles in Europe PMC (number per year are indicated by the blue-colored bars). The author's preprints that have been indexed by Europe PMC are indicated by yellow bars. In 2018, the author's preprints represent 5 of the 15 literature records indexed.

Note that for a publication to be displayed in this profile it must be linked in ORCID as visible to the Public or to Everyone or to Trusted parties, to be displayed. All authors who have an ORCID iD will have a Europe PMC profile page by default. All data displayed on Europe PMC author profiles is openly available from Europe PMC and ORCID.org.

Preprint version management:

The functionality provided thus far has focused on clarifying the "precursor" relationship of any indexed preprint relative to a subsequent publication that is deemed acceptable after peer review. However, a preprint may be the first of a succession of preprint versions that have been revised (e.g. to alter data that is presented and accordingly include/exclude authors) in preparation of the final peer reviewed article version. This calls for preprint version management that would enable for e.g. Europe PMC users to know whether a preprint has other versions, which version they are viewing, which version they are linking to their ORCID profile, or citing. Importantly, linking the preprint versions to each other and to a final "published" version also allows for multiple entry points into the literature graph, i.e. the links to data, ORCID iDs, grant information, citations, that Europe PMC assembles around each literature record.

For the current deliverable, development is being undertaken to make preprint versioning information available for the users of Europe PMC in the following ways:

Firstly, by exposing the versioning information in Europe PMC's Articles restful API, users/developers can gain programmatic access to the various versions of a literature record and associated data.

Secondly, in Europe PMC's website interface, end-users will be able to view version information for an article record on its landing page, as well as in the search results:

- Search results: the "version history" will clarify for an end-user that there are other versions available for the record, and will indicate how many versions are available via "Preprint v1 (or v2)" etc.
- For individual landing pages, the end-user will see hyperlinks to other versions.

Exposure of preprints versioning information to users of Europe PMC requires several items to be prepared. These are itemized in the list below and have been prepared in the order denoted:

1. A layer added to Europe PMC's article database that will house the versioning data

2. Changes to the webservice (that allows access to the database and for its information to be exposed externally via an API response from the Articles restful API)

3. Changes to the search results and to landing pages on the website for individual preprints to reflect the versioning information.

Status of the build

Building in preprint versioning information has been underway at Europe PMC in 2019. In April 2019, steps 1 and 2 above had been achieved and programmatic access to versioning information via Europe PMC's Articles restful API was released. An example of the XML field containing the preprint versioning information that is provided by the core response of the REST search methods can be seen in Figure 23.

For integration of this versioning information into the website interface of EuropePMC: designs have been drawn up of the proposed changes to web pages and examples are shown in Figure 24 and Figure 25. Implementation of these designs began in May 2019.

versionList: states all preprint version list including externalId, source, firstPublicationDate and version number:


```
<versionList>
  <version>
    <id>PPR41846</id>
    <source>PPR</source>
    <firstPublishDate>2018-05-02</firstPublishDate>
    <versionNumber>1</versionNumber>
  </version>
  <version>
    <id>PPR41847</id>
    <source>PPR</source>
    <firstPublishDate>2018-05-23</firstPublishDate>
    <versionNumber>2</versionNumber>
  </version>
</versionList>
```


Figure 23 This figure shows the response retrieved when using the Articles RESTful API to search for the preprint DOI: 10.12688/f1000research.14776.2 (PPR41847). This record represents the second preprint version of the F1000Res preprint entitled: 'Automation of ReactomeFIViz via CyREST API', which was posted in F1000Res on May 23, 2018. What you see here is the XML field in the core response of the REST search methods in the Articles restful web service of Europe PMC, containing the preprint versions available for this preprint and the date on which they were posted. Thus <versionNumber>1</versionNumber> = the first version, PPR41846 posted on 2018-05-02; <versionNumber>2</versionNumber> = the second version, PPR41847 posted on 2018-05-23.

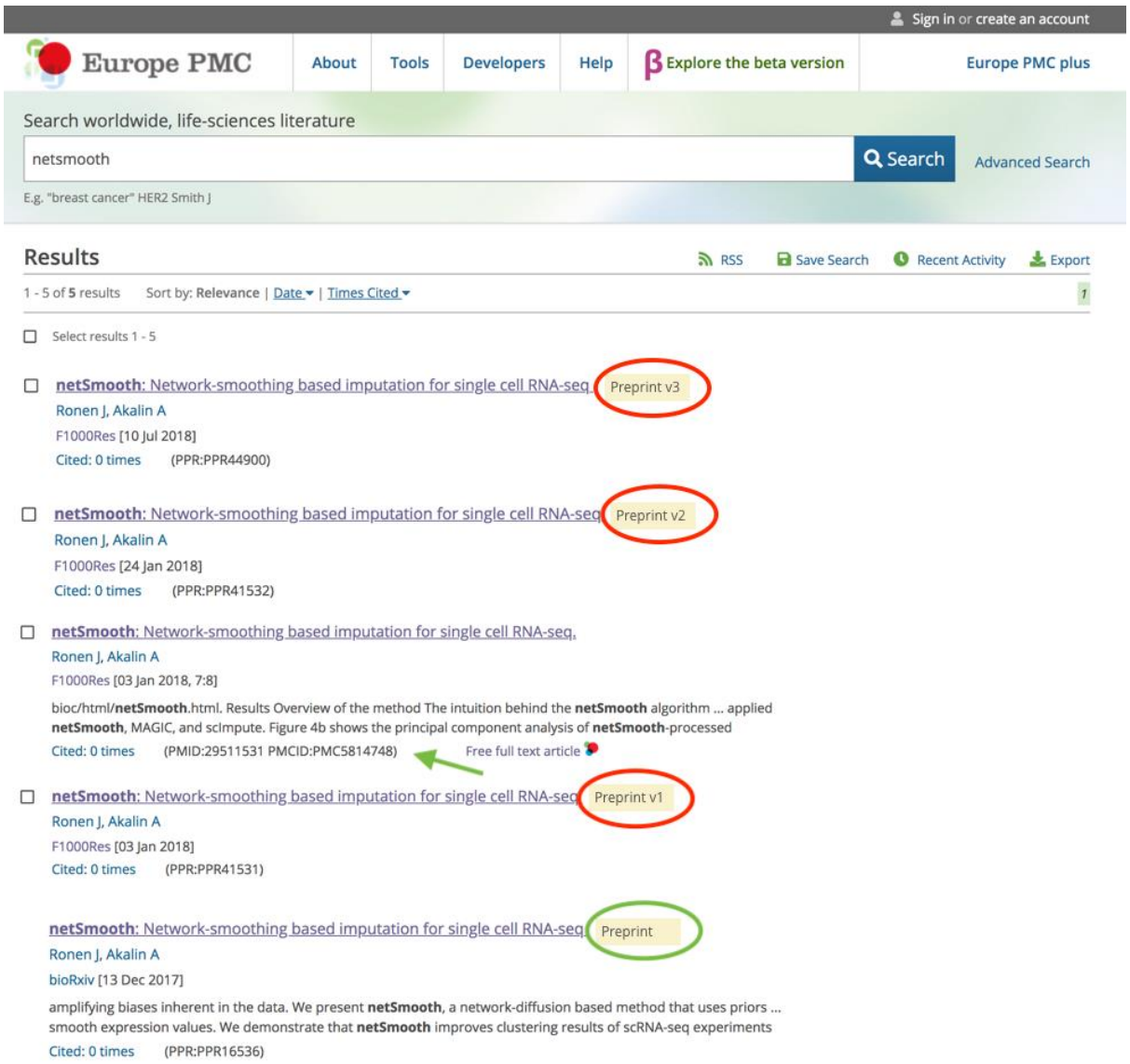

Figure 24 Proposed user interface design (using Sketch) of a search result for a record that has preprint versions denoted by the yellow box indicating 'preprint' that are listed on BioRXiv and F1000Res preprint servers (circled). Also listed is the definitive version that has passed peer review at F1000Research and is assigned the identifiers, PMID:29511531 and PMCID:PMC5814748 (green arrow). Note that the bioRxiv preprint does not have a version number (circled in green). The reason for this is that some preprint servers (including bioRxiv) do not provide version information via Crossref. They only supply the most recent version of the preprint. Therefore Europe PMC can only index one preprint version per record from these preprint servers.

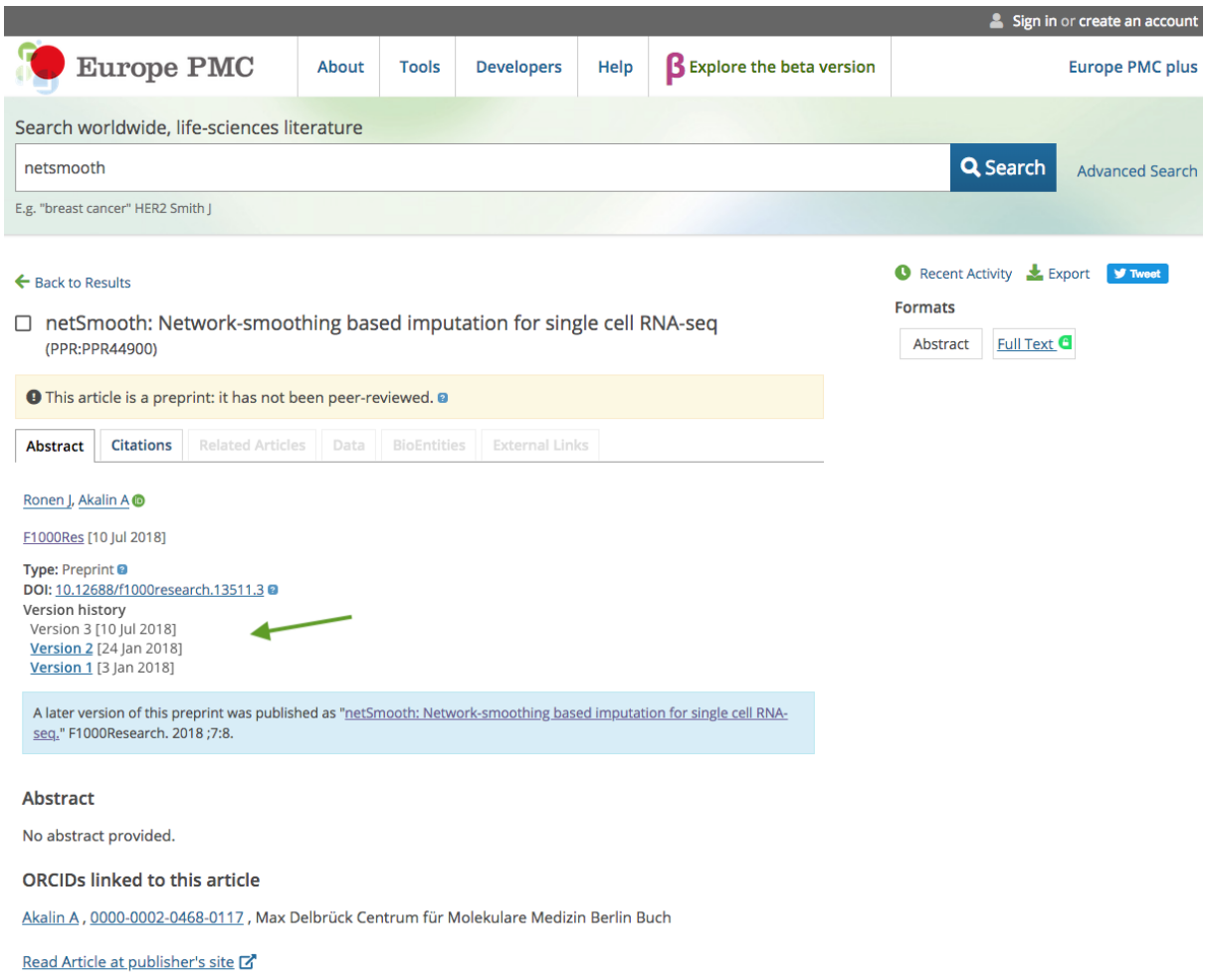

Figure 25 Proposed user interface design (using Sketch) showing the results page for a single preprint record. The version information is displayed for the user, including hyperlinks to other preprint versions where they exist.

Summary table

Pilot Application	**Preprint version management:** Exposure of preprints versioning information to users of Europe PMC - enabling users to see whether a preprint has multiple versions, access those linked versions and where it exists, a final "published" version that has been accepted after journal-organized peer review.
Advanced PID graph functionality	Linking multiple versions of preprints to each other and to the resulting journal article, allows multiple entry points into the knowledge base layers that make up the Europe PMC PID graph.
Method	Europe PMC's article database was extended to include a layer for the versioning data. Changes to the web service enabled access to the database and exposure for users via the Articles restful API. Changes to the search results and content landing pages on Europe PMC's website were made to reflect the versioning information. Note this was applied to preprints from a limited set of providers in this first instance: preprints.org; F1000 Research; all Open Research platforms.

Rationale	Linking preprint versions together provides user access to information that is valuable because: 1.It exposes direct relationships between existing preprint PIDs and journal article PIDs and thereby exposes provenance information for the Journal article 2. it offers transparency into how a published journal article evolves from a preprint that is revised through a series of initial preprint versions. 3. it offers accuracy in being able to answer questions about the time taken from initial preprint version to ultimate publication as a journal article. 4. This pilot (applied to preprints from a limited number of preprint providers) has been built into Europe PMC which is a TRL-9 service housed at EMBL-EBI. The preprint versioning has been flight proven for this initial selection of preprints and along with Europe PMC is a freely available resource for users within EOSC and beyond. The intention is to extend versioning management to include in due course other preprint providers that register version DOI. Version management strategies are important to other databases within the EOSC and lessons learned can be shared.
Target USER	Bioinformaticians, bio-curators who access Europe PMC's web service programmatically; life science researchers & research managers such as librarians who access the database via the web pages.
EOSC	Exposing versioning information in Europe PMC contributes more information for assimilation into the EOSC. Europe PMC's articles restful API and web pages (that include versioning management functionality) can contribute to EOSC services for researchers envisaged by the EOSC pilot partners, including "Citation, attribution, reward", "Peer Review", "publishing support", and "Search & Browse" categories.

3.5 Use case: PID Graph for PhD research supported by large-scale facilities

FREYA Partner: **STFC**

STFC operates several large-scale research facilities such as ISIS neutron and muon facility and Central Laser Facility, and co-owns the Diamond Light Source. PhD students from the UK and across the globe are frequent visitors to the STFC facilities; for the Central Laser Facility, as an example, the share of visiting PhD students constitutes up to 70 per cent of all visitor scientists. Apart from providing grants-in-kind in the form of beamtime awarded to the PhD students, STFC facilities also occasionally play a role of a direct monetary funder of a PhD research, i.e. co-funding certain PhD studentships. Furthermore, about 200 of PhD students in nuclear physics, particle physics and astronomy are funded annually from the block grants that STFC awards to their respective universities.

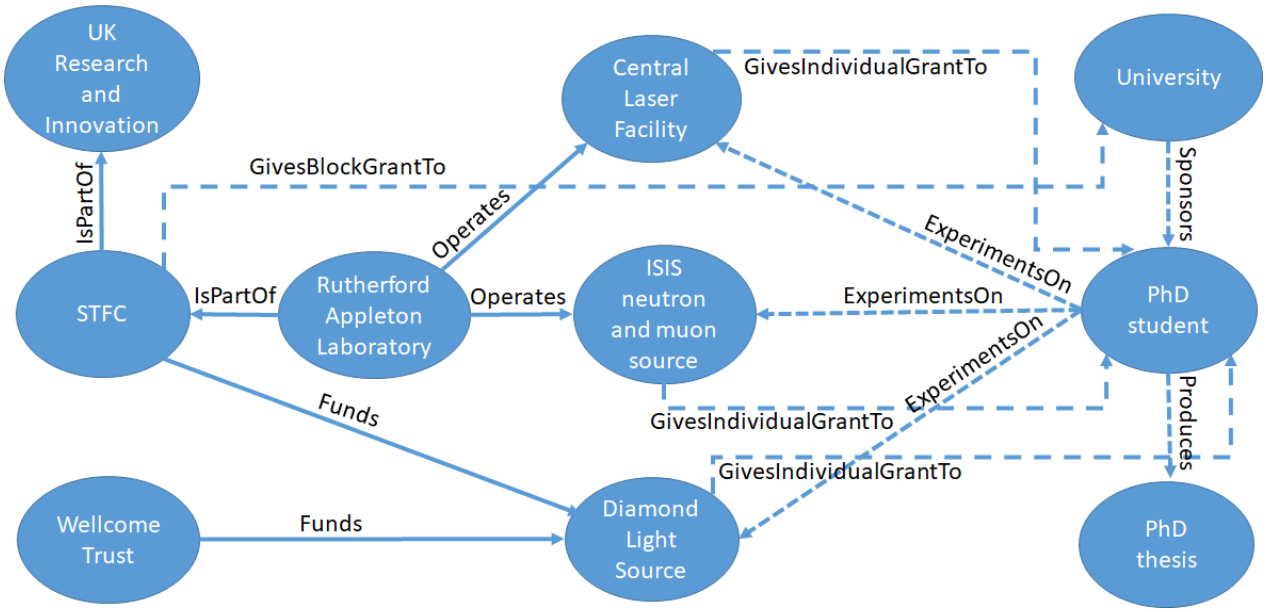

Figure 26. Organizational, operational and funding context of the PhD research supported by STFC.

STFC large-scale facilities and STFC as a whole are interested in measuring research impact of their operation and funding arrangements, including an impact on PhD studies. In FREYA, three user stories collected during activities in WP3 (#35, #68 and #69) were identified as related to the PhD research case. Having abundant PIDs for all elements of PhD research discourse would greatly help to develop the identified user stories, yet there was no out-of-box solution that would allow to measure the proliferation of relevant PIDs in PhD research sponsored by STFC. This is why, a pilot application was required to reflect on the landscape of the PhD research sponsored by STFC, and to take a practical look at the existing and coming PID types that can support impact studies.

The pilot application is a graph database on neo4j graph database platform that used the following data sources:

- Data dump from EThOS system supported by the British Library
- Data dump from GRID.AC supported by Digital Science
- Data extract from STFC ePubs institutional repository
- Data extract from the Diamond Light Source publications database
- Data extract from the research archive of the University of Oxford
- Data extract from Researchfish (the UK Research Information Management platform)

- Data extract from Chemspider supported by the Royal Society of Chemistry

As a first step of data integration, bibliographic records from different sources were compared, with corresponding records discovered and linked. As the next step, certain metadata elements such as the PhD theses authors and their universities have been transformed in separate graph nodes and enriched with additional information from corresponding bibliographic records for authors and from GRID.AC records for university nodes. Then Chemspider nodes were created and connected to selected theses nodes.

The graph allows to answer the following questions:

- Which universities are frequent suppliers of PhD researchers to which STFC facilities?
- (When the source records capture it) Which facility instruments (beamlines) did the prospective PhD experiment on?
- What is the potential for the enrichment of STFC institutional repositories with links to EThOS and other data sources used?
- Which EThOS records currently lack STFC funding information but can be supplied with it using simple inference through the graph connections?
- What are the levels of proliferation of different PID types in the STFC-sponsored PhD research?
- What theses have connections to data in the external reference data sources such as Chemspider?

The graph allows visual reasoning over the involvement of PhD researchers with STFC facilities.

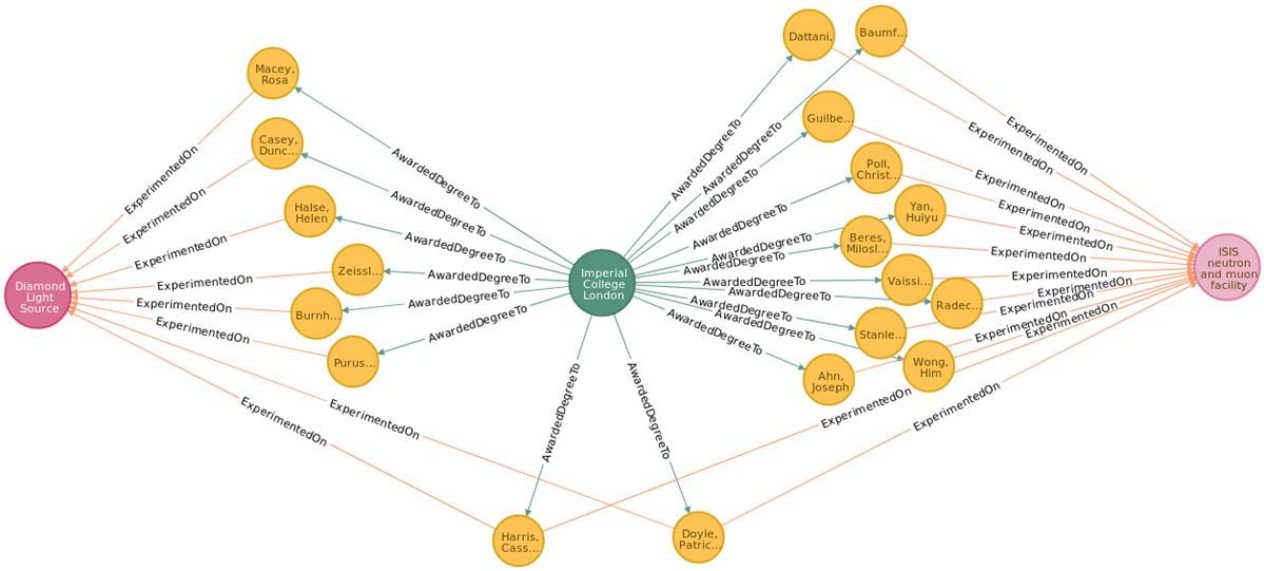

Figure 27. PhD students from Imperial College London who experimented on two STFC facilities. Some of them experimented on Diamond Light Source, some on ISIS neutron and muon source, and some on both facilities.

The PID graph, despite being designed as focused on the STFC information needs, can be used for research discovery or impact studies by third parties, too—as an example, by universities who are frequent suppliers of PhD students in facilities experiments. The graph allows to explore a potential for novel services around PIDs, such as cross-walks between different PIDs for the same entity (Person/Author that can be assigned with ORCID or ISNI, or both, is a good example of how these cross-walks can work).

Also combining information from different sources in the same graph can support better metadata for certain PIDs. As an example, variants of name spellings have been recorded in the graph as AlsoKnownAs property of the Person/Author node as a result of simple inference (collecting the

Author's information from the related graph nodes); this can be used for the extended search and extended exposure of the ORCIDs and ISNIs assigned to the same Person/Author node.

The resulting PID Graph contains nodes and relations that are not particularly specific to the case of a PhD research. STFC is reaching out to Imperial College London that is not a FREYA partner, but is interested in connecting their publications and data repositories to experiments performed on STFC research facilities; these connections can be expressed using the same nodes and relations that are already present in the graph built around the PhD research case.

As a result of collaboration with FREYA, Imperial College London are now introducing a new best practice of minting DOIs for all their doctoral theses, supplied with DataCite metadata that includes organizational identifiers (RORs). Also, they produced examples of data records supplied with DOIs that are connected to thesis records with DOIs; the further intention is to have data placeholders with DOIs for all ongoing PhD research in the Department of Chemistry.

Discussions and initial data exchange with the Cambridge Crystallographic Data Center started, aimed at making connections between facilities research and CCDC records. The case of connecting theses with external reference databases such as Chemspider or Cambridge Crystallographic Data Centre is important, as this gives access to quality data records assigned with persistent identifiers (Identifiers.Org in case of Chemspider and DataCite DOIs in case of CCDC). This work will be continued, specifically in view of a possible route to sustain its outcomes via collaboration with the Physical Sciences Data-science Service.[13]

The lessons learned from building the knowledge graph:

- PhD research is a special case of research supported by STFC as a funder and as a facilities operator, but an interesting and representative one that allows to measure the potential for new information services underpinned by the PIDs and graph database technology.
- PIDs proliferation is substantial for certain information entities such as organizations (with GRID.AC expected to become the main data source for the forthcoming ROR service), patchy for other information entities such as Authors records, and non-existent for information entities such as facility instruments (beamlines).
- Graph database technology is good for information integration from disparate data sources and for the extension with contributions from various third parties.
- The PID Graph that has emerged in support of particular user stories can in fact support other use cases, too, and hence can be considered a seed for a new Open Science infrastructure with a wide range of possible uses.

The PID Graph built around PhD research use case is currently deployed on a virtual machine in a controlled environment and has a standard command line and graphical user interface that is offered by a Community Edition of the neo4j graph database. STFC plans sharing the graph in a more open and sustainable way; the graph can also be exported and shared as a file in a reasonable reusable format such as GraphML that can support a range of tools for the graph exploration and visualization. One natural connection of this pilot to the effort of other FREYA partners is collaboration on the same technology for user interfaces (GraphQL) Another connection is using RORs as PIDs for organizations in addition to GRID.AC PIDs already implemented. These works will be continued within FREYA WP4 and in part WP3 when it relates to new PID types and provenance.

[13] Physical Sciences Data-science Service https://www.psds.ac.uk/

Summary table

Pilot Application	PhD research graph
Advanced PID graph functionality	The graph allows reasoning over PhD research supported by large scale-facilities, using connections among publications, data, individuals and organizations.
Method	Data integration from different sources having different metadata formats. PIDs serve as facilitators of such integration, also in some cases can be propagated through the graph relations to entities that did not initially bear PIDs.
Rationale	The pilot has value for research communities in large-scale facilities and universities with PhD programmes, as it provides a novel research information infrastructure for them with rich context for data and publications. The pilot allows to validate new PID types (RORs for organizations) and technology for PID graphs sharing (GraphQL). It can also be a source of information for making matches between different PIDs for the same entity (e.g. ORCID and ISNI for the same researcher).
Target USER	Librarians, repository managers and research impact teams in large facilities and universities. Researchers (visitor scientists) on large-scale facilities.
EOSC	The pilot can contribute to the EOSC “repository” and “search and browse” service categories. Through the standardized user interface and intensive use of PIDs, it can contribute to other EOSC services (be an information source for them).

3.6 Use case: DANS pilot application

FREYA Partner: **DANS (Data Archiving and Networked Services)**

NARCIS PID-Graph

DANS was not participating is this task, but is asked to contribute to this deliverable. Because DANS didn't develop a particular pilot within this task, the decision is made to report about the NARCIS PID-Graph as a pilot application.

NARCIS is a metadata aggregation for all Dutch Research Information. NARCIS contains metadata from all Dutch Institutional Repositories with publication metadata, all datasets, metadata on research projects, researchers and organizations. All Dutch research organizations and universities are participating in NARCIS.

NARCIS aggregates information on publications and datasets from more than 40 different Dutch repositories and archiving systems and contains only metadata. In addition to the aggregated information from these repositories, NARCIS contains information about all Dutch research organization, researchers and research projects. NARCIS contains many Persistent Identifiers (PIDs). For the different types of publications and datasets DOIs (620.000), Handles (92.000) or URNs (1.8 million). Researcher do also have PIDs: ORCIDs (12.000), ISNIs (26.000) or DAIs (31.000). In FREYA WP4.4, DANS will start to implement organizations PIDs, like ROR, GRID and ISNI. NARCIS also contains projects, although related to researchers and organization, they are not related by PIDs but local IDs.

In many cases, the relations between these objects are known, but frequently the relations between objects are not known and not computer readable. To ascertain these unknown relations, DANS wrote a micro-service that is able to look to other information- and PID services and harvest all PIDs and their relations. In a second step, these relations are mapped against the PIDs that are already in NARCIS and new relations are added. In this pilot, the microservice was configured to look to ORCID.org on the basis of the 12.000 ORCIDs available in NARCIS. All relations attached to these ORCIDs are harvested and added to the NARCIS PID-Graph. The results are very encouraging and DANS plans to harvest other information- and PID services in the near future to expand the NARCIS PID-Graph. Suitable candidates are Dataverse (event) API, Crossref, OpenAIRE funding information and local PURE-instances.

Besides adding more information services, also other PID types like researcher PIDs, organization PIDs, and Funding PIDs can be added to the NARCIS PID-Graph. In the present pilot, only relations were added, but 'new objects' could be added also. Figure 28, 29 and 30 show an example of the effects of the NARCIS PID-Graph. Figure 28 shows the present record of professor M.P. Schijven. Only 2 publications are attached to her name, originating from two different repositories: Utrecht University and Erasmus University. However, NARCIS does contain far more publication for professor M.P. Schijven, but without a person PID.

Figure 28: Example of Prof. Schijven before enriching with the NARCIS PID-Graph. Only 2 publications are related

The next figure (29) shows the situation after a creating a NARCIS PID-Graph with PIDs from ORCID.org. Now 94 publications are added to her name from 6 different repositories: 76 from the University of Amsterdam, 3 from VU University Amsterdam, 3 from Twente University, 1 Radboud University. This is an example of how the NARCIS PID-Graph will relate work to an author. The NARCIS PID-graph will also relate authors to co-authors or data to publications and visa versa.

Figure 29: Example of Prof. Schijven after enriching with the NARCIS PID-Graph. 94 publications are related.

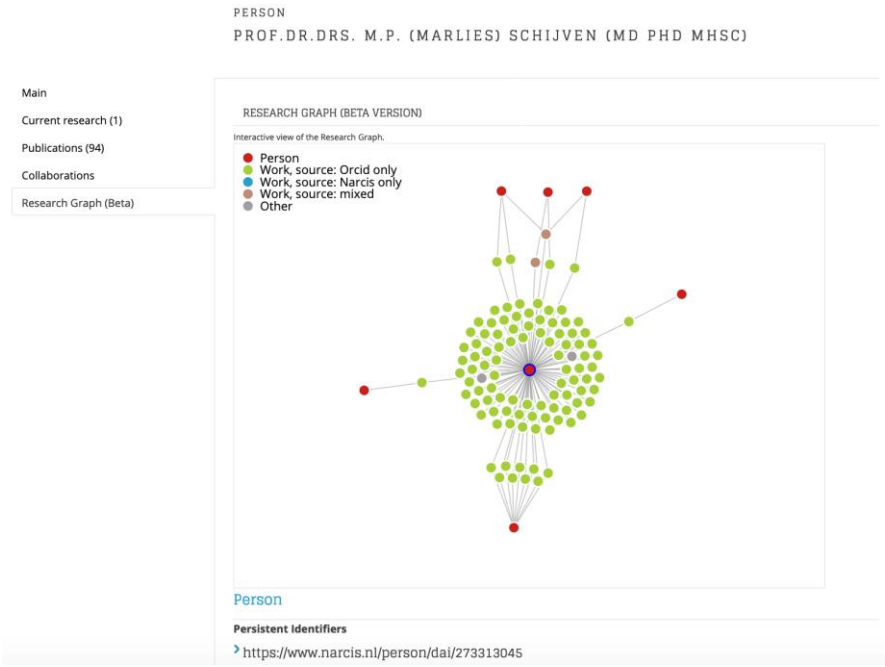

Figure 30 Visualization of the origin of all these relations, together with collaboration network of co-authors.

Summary table

Pilot Application	The NARCIS PID-Graph. New functionality added to the National Academic Research and Collaborations Information System (NARCIS) in WP4.
Advanced PID graph functionality	PID-Graphs from other services are added to the NARCIS PID-Graph, and so enrich relations between digital objects in NARCIS. The content contains information from more than 40 institutional repositories, the Dutch Research Council and other sources. Relations between these objects are enriched by ORCID.org, which contains information from sources like DataCite, CrossRef. The Graphs are made available again through Linked Data.
Method	DANS wrote a microservice to 'harvest' relations from other information services to create a NARCIS PID-Graph. NARCIS offers these relations in Linked Data (JSON, Schema.org) to other participants.
Rationale	The purpose of this application is to enrich Dutch Research information by using relation between PIDs from other services in the Dutch Research Information Infrastructure. Information of scientific output is often scattered, data is archived over many services and the scientific community and funders do not have the necessary overviews. This application gather Research Information from different sources and connects it to the personal page of the researcher.
Target USER	a) End-users of the NARCIS web interface, who will see research more in context because the digital objects are more related to each other. b) Stakeholder of NARCIS and repository managers of Dutch research organizations, who can add enrichment to their own system. Code is available op GitHub https://github.com/DANS-KNAW/narcis-pid-graph https://github.com/DANS-KNAW/narcis-pid-aggregator https://github.com/DANS-KNAW/narcis-pid-manager
EOSC	The information of the NARCIS PID-Graph is available for all other services. NARCIS is already connected to Open-AIRE and Be2Share portal. NARCIS is also a 'bridge' to OpenAIRE for all Dutch research organizations. Through open standards interoperability with other services in the EOSC is assured.

4 Conclusions

The work leading to this deliverable challenged the FREYA partners to advance their PID Graph functionality in a disciplinary context, and thereby provide improved user access and discoverability of their resources. This was accomplished through the development of new tools, enabling users to better navigate the disciplinary PID Graph of the respective partners. The work on advancing PID Graph functionality takes its starting point in use cases expressing a specific user need, which can be accommodated though a better way to navigate the PID Graph. The use cases presented in this deliverable were either collected during activities in WP3 (D3.2) or express a use case specific to the individual partner.

A diverse set of "PID Graph navigation" tools were built, which illustrate the great heterogeneity within the FREYA consortium and the variety of challenges needed to solve for the respective communities:

- **PANGAEA** focused on providing a PID Graph with a user interface to easily search a sample's research environment, i.e. related samples, datasets and publications, and scientists with the mature PIDs available for these entities

- **BL** and **EBI** provide tools that consolidate identical documents or identify different versions of the same document in a format that can be easily understood by users. Following a general user need in regards to managing duplication and versioning of research output or other items.

- **CERN** focused their work on automating workflows to enrich metadata and enhance their CAP-PID Graph by connecting content from multiple sources to CAP for a more seamless user experience. In addition, they have implemented an integration of Python Notebooks directly in the user interface of CAP, so that users can explore the output of their analysis live through a visualization tool.

- **STFC** provided a tool for universities and research infrastructures to track their impact in graduate education by focusing on the activities of PhD student using STFC facilities. The STFC graph is focused on activities of PhD students, but the connections expressed through nodes and relationships in the current graph can be used for other applications as well and interest in this has already been stated by the Imperial College London.

The PID Graph advancements, described by the individual partners, are at different levels of readiness; some of the tools presented are fully developed and will be implemented and available to users in a matter of weeks, while others are presented as blueprints awaiting production at a later stage.

Common for the tools developed is that they will be directly or conceptually useful for other data holders from other disciplines. The transfer of graph functionality to new services, data providers/repositories or users is also essential to the sustainability of FREYA output. For this purpose, the tools will be registered with the EOSC service registry, allowing the new PID Graph functionality and the tools facilitating it to be discovered, adapted, and implemented by others for the same or extended purposes. This implementation will be fulfilled once each of the services has research the necessary maturity level required by EOSC.

FREYA's "BoF" initiative at the 14th RDA Plenary in Philadelphia resulted in ongoing efforts to form a Research Graph Interest Group, that will investigate if and how the different graph systems inside and outside of FREYA (e.g. OpenAIRE, Research Graph and the many others) can be harmonized, centralized or adapted to a degree where they run complementary and actively support Open and FAIR Science in the European and global landscape. Consolidating and adapting the FREYA graph to the current graph landscape is essential to ensure the external adoption and sustainability of the project output.

5 EOSC perspective

The pilot applications developed in FREYA contribute to many of the EOSC services specified in the EOSC service architecture[14]. The applications are particularly relevant for the "PID handler", which is defined as a class of services to generate and resolve Persistent Identifiers (PIDs). EOSC recommends assignment of PIDs as a way to make the resources available in accordance with the FAIR principles (making resources findable, accessible, interoperable and reusable). The Pilot applications demonstrated in this deliverable go beyond mere resolution and provide services and tools developed to exploit the existing PID landscape, providing easy access to links to other relevant resources using the PID graph. The developed services enable users to find specific resources using PIDs and also to resolve PIDs to provide access to the resources themselves.

The pilot applications also contribute to many of the services for researchers envisioned in the EOSC services architecture such as:

- **The "search and browse",** which is a class of services to enable researchers to find research artefacts (e.g. datasets, papers, methods, workflows) matching their interests. The pilot applications demonstrated in this deliverable all revolves around equipping the user with better ways to find relevant research resources and to distinguish different versions of an artefact using PID graphs. This contributes to a better search and browse functionality.
- **The "Recommender",** which is a class of services envisaged to provide researchers with objects of potential interest by predicting their preferences. The pilot applications introduced in this deliverable all revolve around improvements of discovery and access to resources. Using the PID graph to exploit other resources linked to a specific resource with a specific PID. In this way, the tools developed also function as a recommender service for finding related resources and presenting (recommending) them to the user.
- **The "Data Taming",** which is a class of services supporting researchers in the development of suitable datasets for their research. This consists of identifying datasets of potential interest and extracting from the dataset items of interest. The pilot applications demonstrated in this deliverable all enable the user to better identify resources of potential interest by exploring the PID graph and thereby contribute to a higher degree of data taming

The pilot applications in this deliverable were developed during the FREYA project and are still in a developmental stage, and as such are not yet ready for implementation in the EOSC services catalog, which is restricted to services with TRL 8 and 9. The catalogue strives to include only services, which are considered mature enough to be at production level. The pilot applications demonstrated here are at the most at a TRL 6, and many are lower. Hence, implementation of these services in the EOSC service catalogue is not yet possible, but implementation is envisaged as soon as the needed TRL is reached. In a recent dialogue between FREYA and EOSC, it was suggested that the criteria for implementation in the EOSC services catalog might be lowered allowing the most mature applications coming out of FREYA to be implemented. FREYA will follow up on this initiative at a later stage in order to evaluable the possibility for future implementation. In general, FREYA is contributing services to the EOSC services catalog, but the pilot applications demonstrated in this deliverable are not included in the first round of implementation.

[14] EOSC-PILOT D5.4: Final EOSC Service Architecture Ref. Ares(2019)2469866 - 08/04/2019