



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](https://orcid.org/0000-0002-5948-0980))  
Ketil Koop-Jakobsen (PANGAEA, [orcid.org/0000-0002-1540-6594](https://orcid.org/0000-0002-1540-6594))  
Uwe Schindler (PANGAEA, [orcid.org/0000-0002-1900-4162](https://orcid.org/0000-0002-1900-4162))  
Artemis Lavasa (CERN, [orcid.org/0000-0001-5633-2459](https://orcid.org/0000-0001-5633-2459))  
Pamfilos Fokianos (CERN, [orcid.org/0000-0003-0618-1722](https://orcid.org/0000-0003-0618-1722))  
Sünje Dallmeier-Tiessen (CERN, [orcid.org/0000-0002-6137-2348](https://orcid.org/0000-0002-6137-2348))  
Vasily Bunakov (STFC, [orcid.org/0000-0003-3467-5690](https://orcid.org/0000-0003-3467-5690))  
Frances Madden (British Library, [orcid.org/0000-0002-5432-6116](https://orcid.org/0000-0002-5432-6116))  
Christine Ferguson (EMBL-EBI, [orcid.org/0000-0002-9317-6819](https://orcid.org/0000-0002-9317-6819))  
**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

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



# FREYA project summary

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

The vision of FREYA is built on three key ideas: the **PID Graph**, **PID Forum** and **PID Commons**. The PID Graph connects and integrates PID systems to create an information map of relationships across PIDs that provides a basis for new services. The PID Forum is a stakeholder community, whose members collectively oversee the development and deployment of new PID types; it will be strongly linked to the Research Data Alliance (RDA). The sustainability of the PID infrastructure resulting from FREYA beyond the lifetime of the project itself is the concern of the PID Commons, defining the roles, responsibilities and structures for good self-governance based on consensual decision-making.

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

For more information, visit [www.project-freya.eu](http://www.project-freya.eu) or email [info@project-freya.eu](mailto:info@project-freya.eu).

---

## Disclaimer

This document represents the views of the authors, and the European Commission is not responsible for any use that may be made of the information it contains.

## Copyright Notice

Copyright © Members of the FREYA Consortium. This work is licensed under the Creative Commons CC-BY License: <https://creativecommons.org/licenses/by/4.0/>.

## 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.

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<sup>1</sup>) 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 revolving 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 the 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.

---

<sup>1</sup> [https://www.project-freya.eu/en/deliverables/freya\\_d3-2.pdf](https://www.project-freya.eu/en/deliverables/freya_d3-2.pdf)

# Contents

1	Introduction.....	5
1.1	PID implementation.....	5
1.2	PID Graph functionality .....	5
1.3	PID Graph functionality in FREYA .....	6
2	Methods .....	7
2.1	Use cases as a methodological approach .....	7
	Advanced PID Graph functionality established .....	8
3	Use cases .....	9
3.1	Use case: Persistent Identifiers for physical samples (IGSNs) .....	9
3.2	Use case: Persistent Identifiers for Documents (DOIs) .....	17
3.3	Use case: CERN Analysis Preservation (CAP)—a community PID Graph.....	24
3.4	Use case: Exposing provenance information around preprints .....	30
3.5	Use case: PID Graph for PhD-student activities and outcomes .....	37
4	Conclusions.....	40

# 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<sup>2</sup>) 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 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 user 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 and simple functions to navigate the network and Open Science landscape. Prior to work done by FREYA partners, the

---

<sup>2</sup> [https://www.project-freya.eu/en/deliverables/freya\\_d3-1.pdf](https://www.project-freya.eu/en/deliverables/freya_d3-1.pdf)

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 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 functionality 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 functionality within their domain. 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 users 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 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 the tools and services developed registered in the EOSC services registry.

## 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 workplan.

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 a development that actually benefits the end-users, and thereby generates uptake of PIDs in the communities. A use case revolving around users of Open Science and that describes an actual scenario, where technological development can improve the user experience. FREYA has previously collected a list of 75+ use cases from the Open Science community<sup>3</sup>. 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) and when the necessary tools to navigate the PID graph are provided and thereby 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.

The prototype applications are at different stages of technological development and this deliverable focuses on the conceptual development of PID Graph functionality. It identifies new PID Graph functionality that will fulfill specific needs for the PID Graph use and providing a blueprint for functionality and 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.

---

<sup>3</sup> <https://www.pidforum.org/c/user-stories>

## Advanced PID Graph functionality established

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, which can extract essential metadata information through the linkages of IGSNs, ORCIDs, and DOIs for articles and data, and make it available to 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 organisations 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 visualisation 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. Theses 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<sup>4</sup> as a precursor for the forthcoming ROR<sup>5</sup> 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.

---

<sup>4</sup> <https://www.grid.ac/>

<sup>5</sup> <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 Reefer in MARUM I building, Bremen Universität © 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.

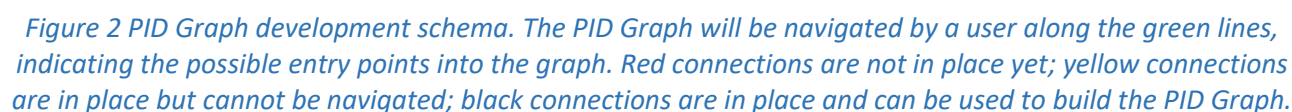
In the previous deliverable, D4.1, we demonstrated how PANGAEA has successfully integrated IGSNs in some of their published datasets, allowing the IGSN to be displayed in the datasets along with other essential persistent identifiers, such as ORCID for authors, DOIs for articles and DOI for the dataset itself. Hereby, we have generated a PID Graph that facilitates access to essential metadata about the sample with an IGSN using persistent identifiers. However, it requires the user to manually “click” their way through the actionable PID-links available and compile the 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 has IGSNs, whereas additional information available from sources outside PANGAEA is 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

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.

For the users, we want to expand the PID graph functionality to encompass the following:

- The APP will explore the following PID graph (Figure 2).



## **Automated compilation**

The APP 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 the APP, we will automate this search, building APIs to enter the PID Graph and automatically compile information starting with related data-set and articles from various sources.

## **Extended metadata collection**

The APP 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. 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 the APP, 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 hierarchal 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 bore hole can contain multiple sediment cores. After curation in the core repository, samples can be taken from the cores and send out to scientists around the world. Individual unique IGSNs are assigned to bore holes, cores and samples. Furthermore, expeditions have their own identifier in form of a number; although not a PID, the expedition numbers are also registered with the dataset 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. The APP 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.

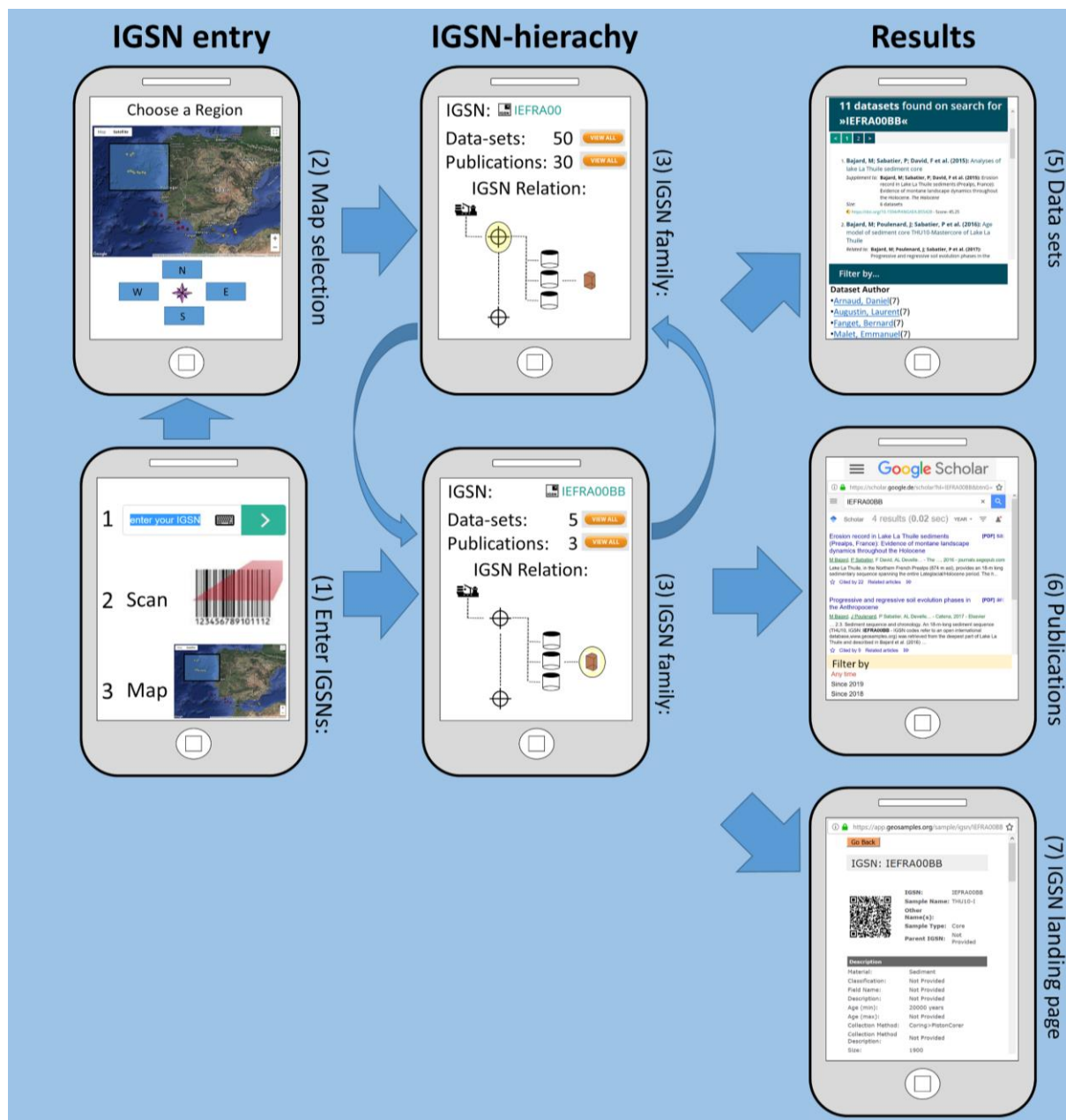
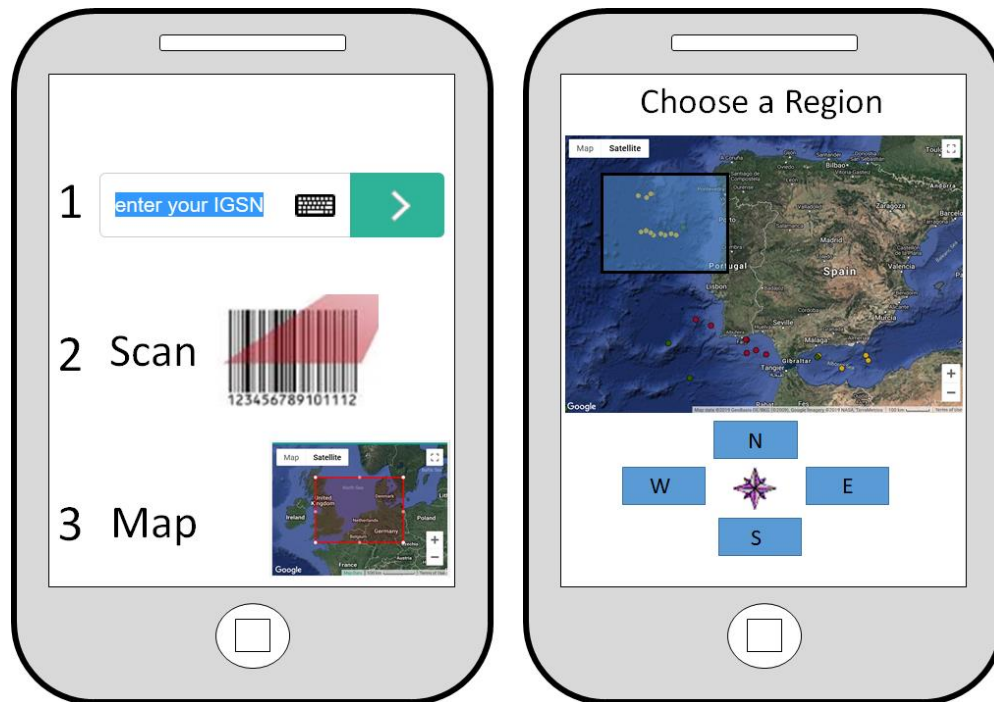


Figure 3 Workflow for the IGSN APP for Bremen Core Repository showing the pathway from IGSN submission to related Data collection, publications collection and corresponding IGSN landing page




## IGSN entry sites

The entry site of the APP (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.

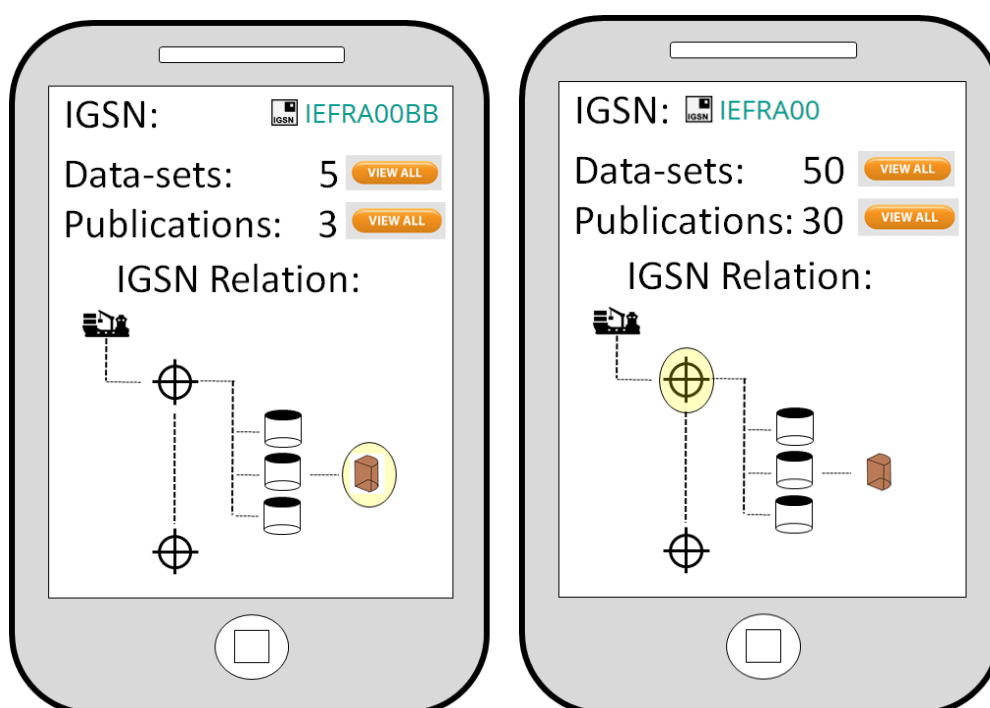






*Figure 4 IGSN Entry sites (Left). Give IGSN by (1) type in the IGSN number (2) Scan a barcode on the cores (3) select IGSNs from map. Map-selection (Right), draw the area in question on the map.*

## 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 authors 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.



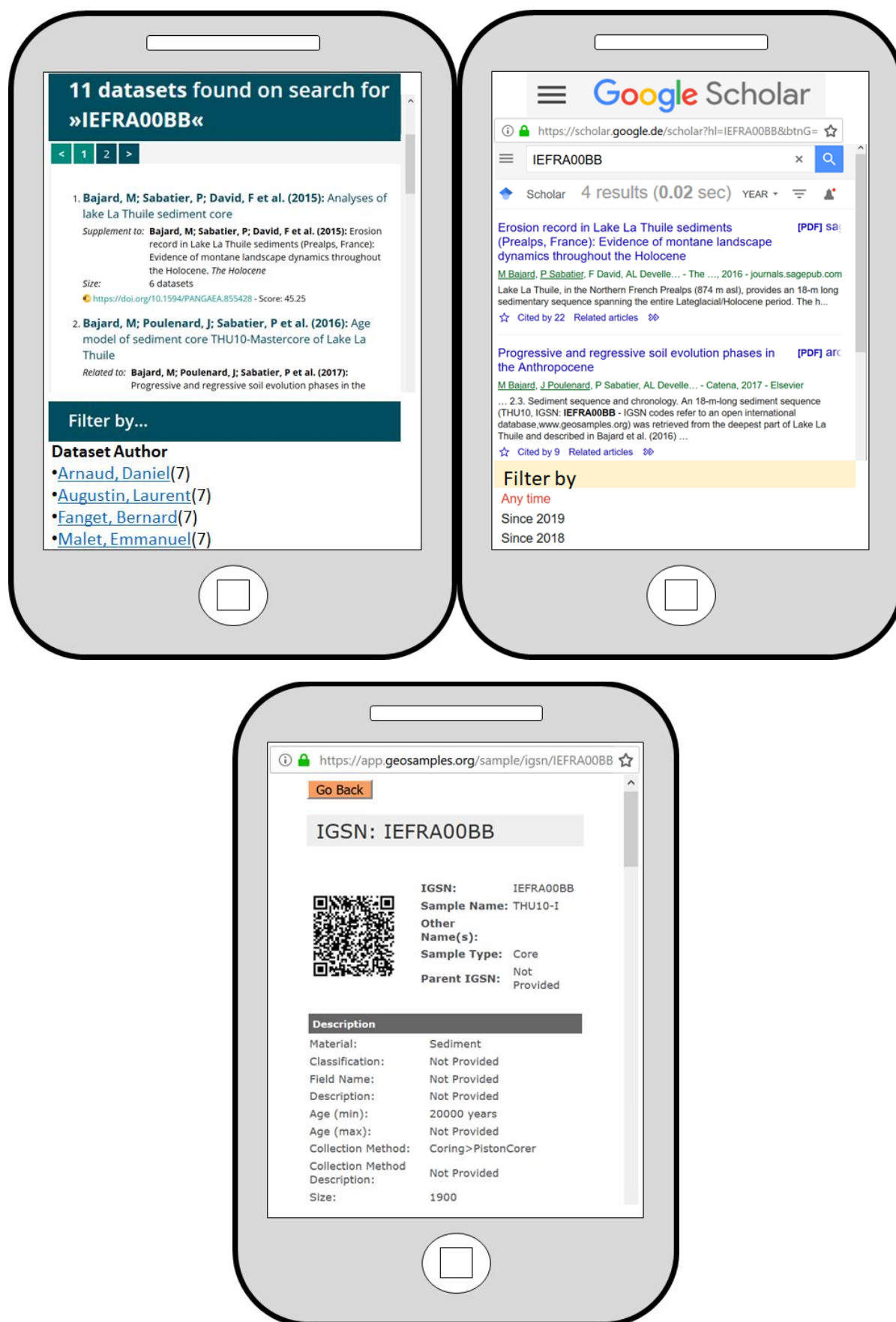


Figure 6 Result sites showing the data set (top-left), publications (top-right) and IGSN-landing page for a given IGSN (bottom)

## Perspective

The APP is meant as a demonstrator showing the benefits of linking IGSN-related PIDs in a PID Graph and developing software that can help the user navigate the PID Graph. The first version of the APP will take a starting point in the IGSNs from sediment cores in the Bremen Core Repository and related metadata stored in the PANGAEA database, as well as metadata stored externally at FREYA partner, DataCite. It is a future goal to expand the reach of the APP, building API that can go beyond PANGAEA and FREYA partners and facilitate global searches for IGSN-related metadata. This includes 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 the APP working on a regular cell phone and in this way be a practical tool that can be used while working on the cores in the laboratory.



## 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<sup>6</sup>, 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.

The screenshot displays the BL's discovery service interface. At the top, there are navigation links: 'Main catalogue', 'Our website', and 'Explore Further'. Below these is a search bar containing the text 'squatters as folk devils' and a dropdown menu set to 'Everything in this catalogue'. To the right of the search bar is a magnifying glass icon and a link to 'Advanced search'. On the left side, there is a sidebar with filters. Under 'Access Options', there are links for 'Request to Reading Room (1)', 'Online: Reading Room only (1)', and 'Purchase a copy (1)', along with a 'Refine further' button. Under 'Subject', there are links for 'Deviant behavior--Periodicals (1)' and 'Deviant behavior (1)', also with a 'Refine further' button. Below the filters is a section 'Try a new search' with 'by this Author/Contributor: Dee, E. T. C.' and 'on this subject: Deviant behavior--Periodicals, Deviant behavior'. The main results area shows '2 results for Everything in this catalogue' sorted by 'relevance'. Both results are identical and list the article 'The Production of Squatters as Folk Devils: Analysis of a Moral Panic that Facilitated the Criminalization of Squatting in the Netherlands' by Dee, E. T. C. from the journal 'Deviant behavior: an interdisciplinary journal -- Taylor & Francis -- Volume: 37 7; (pages 784-794) -- 2016'. Each result has a 'Details' link and an 'I want this' link.

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.<sup>7</sup> In Figure 9 only the access options for Legal Deposit<sup>8</sup> content are displayed because it uses the Legal Deposit Archive Resource Key (ARK) to display this access route.

<sup>6</sup> <http://explore.bl.uk>

<sup>7</sup> The BL's On Demand Service provides copies of books and articles for a fee. <https://www.bl.uk/on-demand>

<sup>8</sup> 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.

## The Production of **Squatters as Folk Devils**: Analysis of a Moral Panic that Facilitated the Criminalization of Squatting in the Netherlands

Dee, E. T. C.

Deviant behavior: an interdisciplinary journal -- Taylor & Francis -- Volume: 37 7; (pages 784-794)


-- 2016

[Details](#)
[I want this](#)

---


Which of the options below can I use? Actions ▼

**Deliver item to**

 My Home or Office [Go](#)

This option allows anyone to order copies. Copyright charges may apply. See our website for more information on [On Demand](#) and [help](#) with ordering. NB: Readers wishing to order to their home or office must also be registered for our On Demand service. You can register via [My Account](#).

**Feedback**

 Problems or Comments? Complete the Feedback Form [Go](#)

NB: You may find our [Help](#) can answer your query.

*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.*

## The Production of **Squatters as Folk Devils**: Analysis of a Moral Panic that Facilitated the Criminalization of Squatting in the Netherlands

Dee, E. T. C.


Deviant behavior: an interdisciplinary journal. Volume 37:Issue 7 (2016); pp 784-794 -- Taylor & Francis

[Details](#)
[I want this](#)

---


Which of the options below can I use? Actions ▼

**View online**

 Legal Deposit content: only available on Reading Room computers [Go](#)


The British Library's [Legal Deposit](#) collections.

**Deliver item to**

 The British Library's Reading Rooms (for registered Readers only) [Go](#)

NB: Only [Readers](#) can view print articles in our Reading Rooms. You will need the article's year, volume and issue details when ordering the title.

**Feedback**

 Problems or Comments? Complete the Feedback Form [Go](#)

NB: You may find our [Help](#) can answer your query.

*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<sup>9</sup>.

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 to

<sup>9</sup> <https://unpaywall.org/>

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.

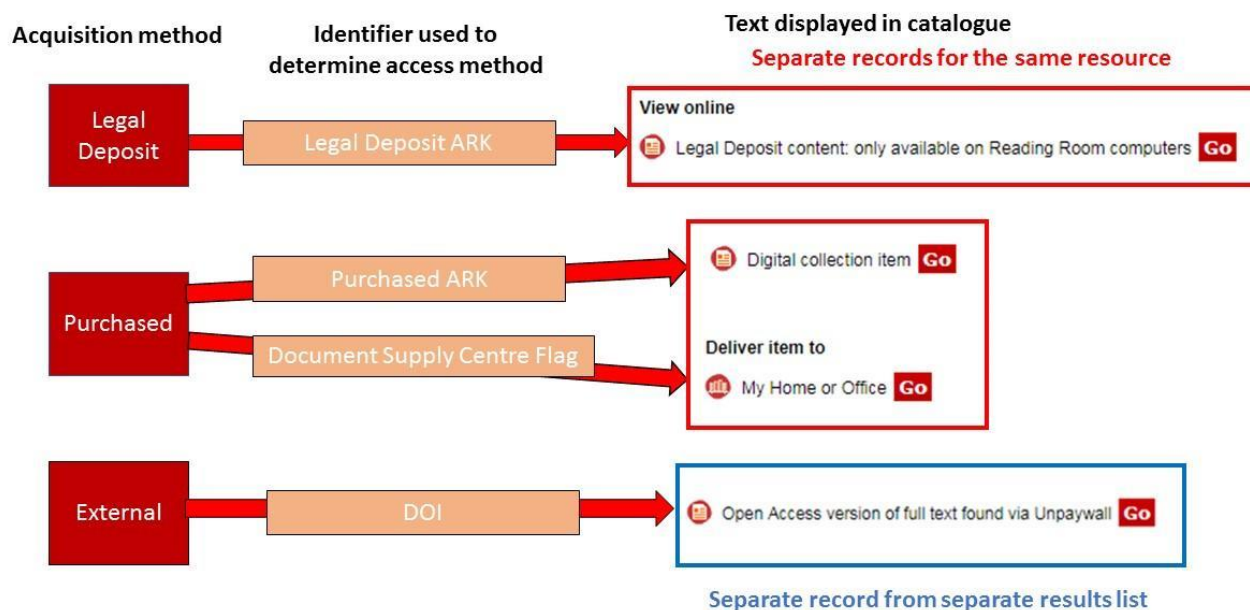


Figure 10 Diagram representing the current discovery experience. The different acquisition routes (left) provide different access text (right), using different identifiers, depending on the acquisition route. The different acquisition methods also lead to the creation of separate records for the same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

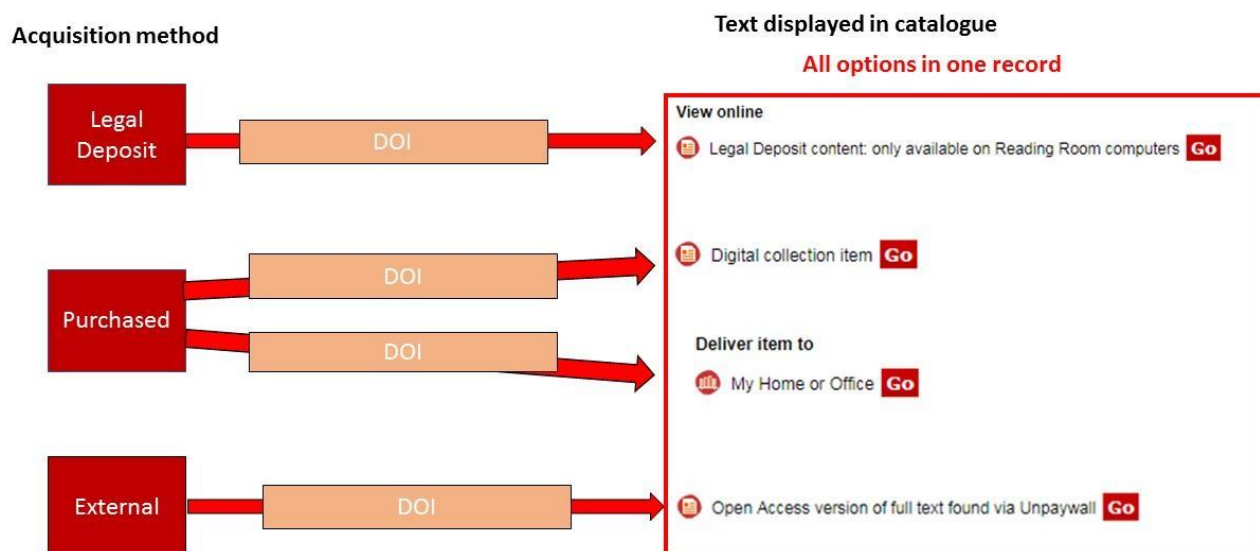


Figure 11 Diagram representing the planned development. All access options will display in one record allowing the user to make an informed choice of access route. The DOI will be used to match the results from the different acquisition routes to one record.

### The Production of **Squatters as Folk Devils**: Analysis of a Moral Panic that Facilitated the Criminalization of Squatting in the Netherlands

Dee, E. T. C.

Deviant behavior: an interdisciplinary journal. Volume 37:Issue 7 (2016); pp 784-794 -- Taylor & Francis

Details I want this

Which of the options below can I use?

Actions ▼

#### View online

- Open Access version of full text found via Unpaywall **Go**
- Legal Deposit content: only available on Reading Room computers **Go**
- Digital collection item **Go**

#### Deliver item to

- The British Library's Reading Rooms (for registered Readers only) **Go**

NB: Only **Readers** can view print articles in our Reading Rooms.  
You will need the article's year, volume and issue details when ordering the title.

- My Home or Office **Go**

Figure 12 A mock up 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 rationalise their access to content. The BLDOI service provides a mechanism to streamline the data integration between the BL and other UK Legal Deposit Libraries.<sup>10</sup> 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.

---

<sup>10</sup> <https://www.bl.uk/legal-deposit/about-legal-deposit>

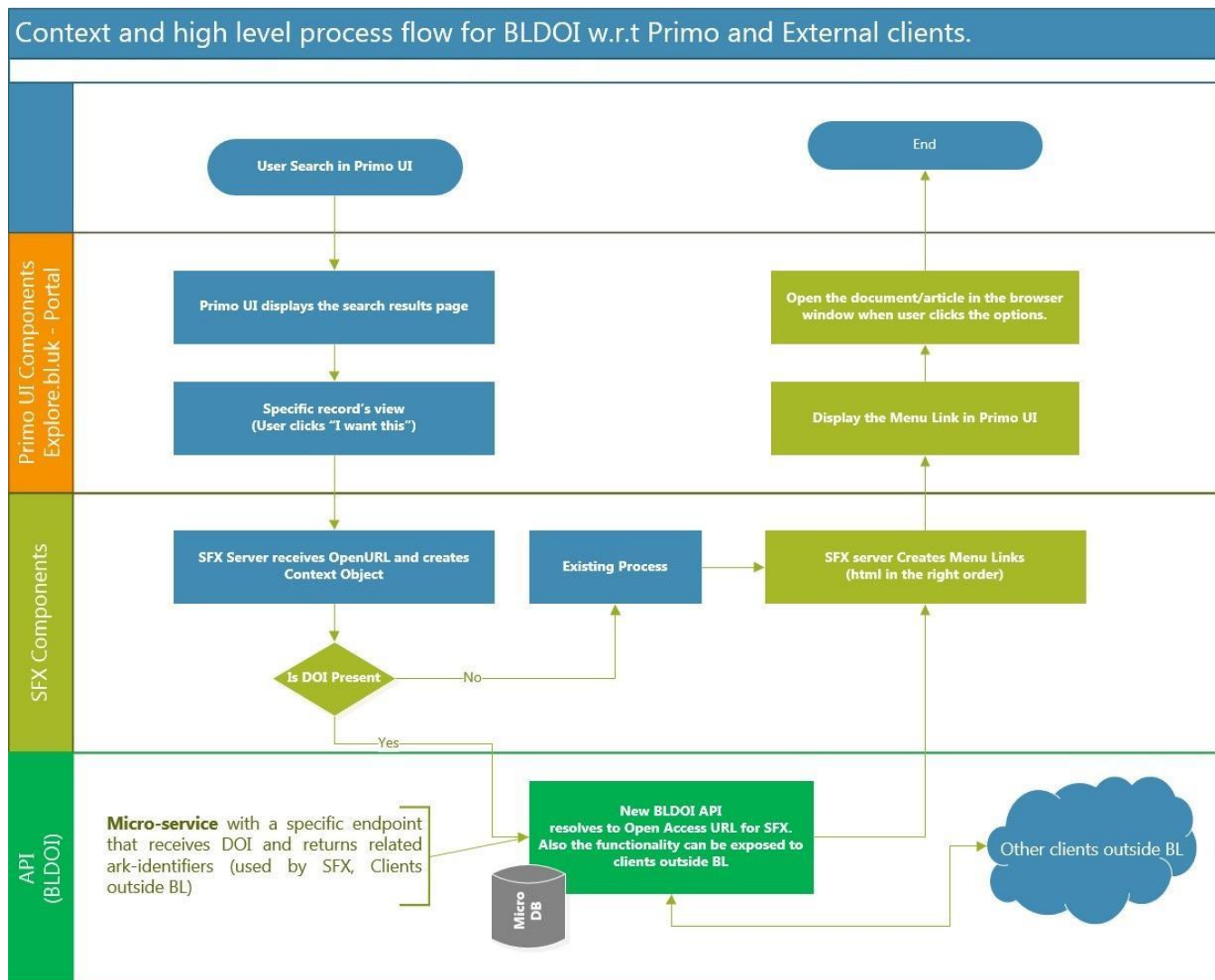


Figure 13 A swimlane diagram demonstrating what the planned developments (bottom lane) will deliver from the point of view of the process from a user's search to the link to the resource being returned to the user to click on. It illustrates how this new service (bottom lane) will interact with the existing 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.



## 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 visualisation 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)<sup>11</sup>, 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 visualisation 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 visualisation 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, such features also enhance findability and interoperability of information, while also enriching the CAP PID Graph.

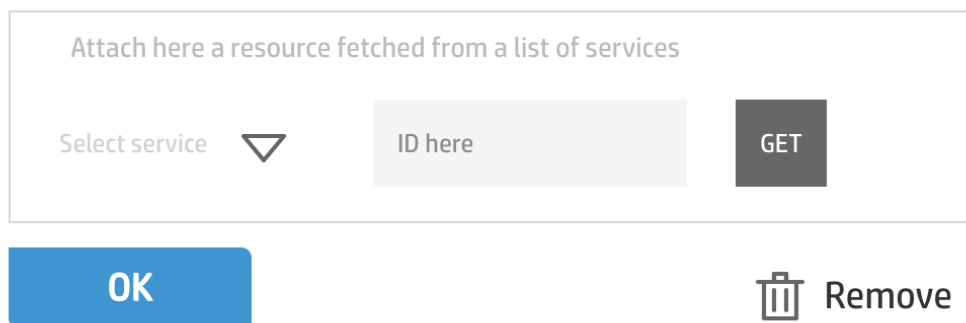
---

<sup>11</sup> <https://github.com/cernanalysispreservation/analysispreservation.cern.ch>



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).

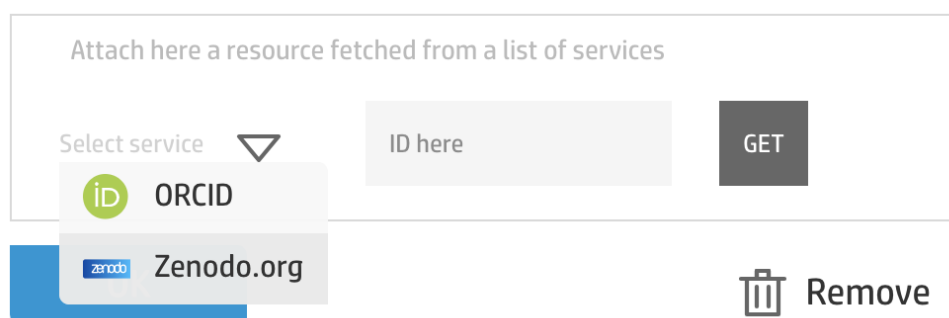


Attach here a resource fetched from a list of services

Select service ▼ ID here GET

OK Remove

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



Attach here a resource fetched from a list of services

Select service ▼ ID here GET



iD ORCID


zenodo Zenodo.org

OK Remove

*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.*



Attach here a resource fetched from a list of services


ID:	0000-0003-0618-1722		
Name:	Fokianos Pamfilos		
URL:	<a href="https://orcid.org/0000-0003-0618-1722">https://orcid.org/0000-0003-0618-1722</a>		

**OK**  Remove

*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).*

Attach here a resource fetched from a list of services

ID:	2644242		
Title:	UNEDLabs/rip-spec: Draft incomplete release		
DOI:	10.5281/zenodo.2644242		
URL:	<a href="https://zenodo.org/api/records/2644242">https://zenodo.org/api/records/2644242</a>		

**OK**  Remove

*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"
  }
  "fetches" : { 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.

## Visualisation 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.

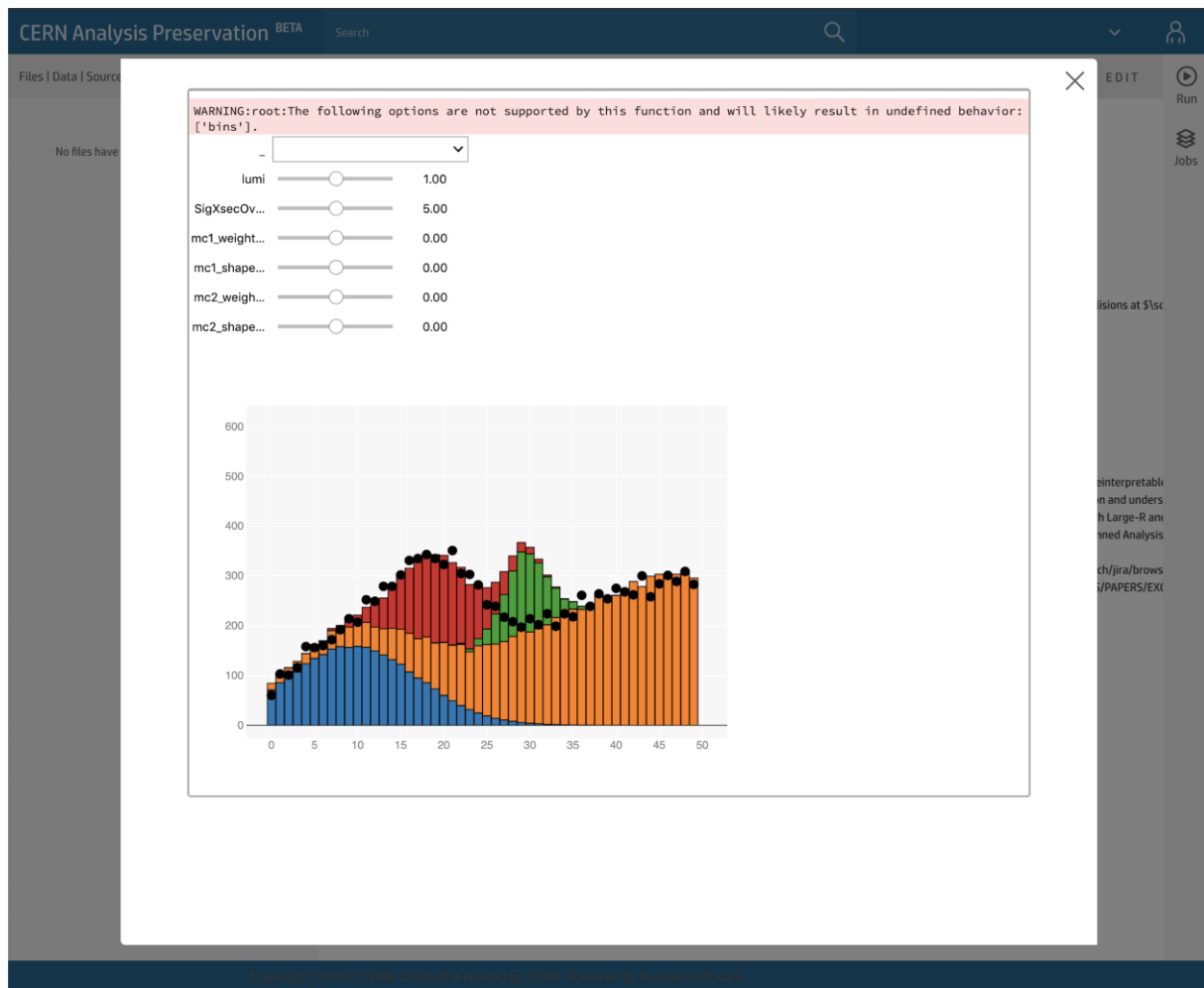
The screenshot displays the CERN Analysis Preservation (CAP) web interface. At the top, there's a header with the text 'CERN Analysis Preservation BETA' and a search icon. Below the header, there are tabs for 'Files | Data | Source Code' and a 'PUBLISHED' status indicator. The main content area is divided into two columns. The left column contains a message: 'No files have been attached to this analysis.' The right column displays detailed information about a specific analysis, including 'BASIC INFORMATION' and 'INFORMATION FROM GLANCE DATABASE'. The 'Click to Visualize' button is highlighted with a black circle.

BASIC INFORMATION	
ANALYSIS TITLE	MET + H->bb search 13 TeV 2017
GLANCE ID	225

INFORMATION FROM GLANCE DATABASE	
GLANCE ID	225
SHORT TITLE	JDM - mono-H(bb) 2017
FULL TITLE	Search for Dark Matter Produced in Association with a Higgs Boson Decaying to $b\bar{b}$ using $36\text{ fb}^{-1}$ of pp collisions at $\sqrt{s}=13\text{ TeV}$
PUBLICATION TITLE	MET + H->bb search 13 TeV 2017
REF CODE	ANA-EXOT-2018-01
CREATION DATE	2018-02-05
STATUS	phase0_closed
PHASE 0	
DATASET USED	2015-2017 with release 21
ID	263
MAIN PHYSICS AIM	Search for new physics in the signature of H->bb and MET. Motivated by Dark Matter models but also want to produce reinterpretable boosted/merged regime and resolved regime. Non-physics goal: Include 2017 data to validate release 21 reconstruction and unders
METHODS	Previous analysis was EXOT-2016-25. * Planned CP Techniques (beyond standard object reconstruction/selection): Both Large-R and corrections for semi-leptonic b-hadron decays, Variable Radius track jet b-tagging, MET significance (object based) Planned Analysis for background estimation from Monte Carlo with several exclusive signal and control region selections.
MODEL TESTED	-Benchmark model used: Z'-2HDM : <a href="https://arxiv.org/abs/1402.7074">https://arxiv.org/abs/1402.7074</a> MC production documented here: <a href="https://its.cern.ch/jira/browse/INDEPENDENT LIMITS ALSO TO BE PROVIDED, SIMILAR TO THE PREVIOUS RESULT: https://atlas.web.cern.ch/Atlas/GROUPS/PHYSICS/PAPERS/EXOT-2016-25">https://its.cern.ch/jira/browse/INDEPENDENT LIMITS ALSO TO BE PROVIDED, SIMILAR TO THE PREVIOUS RESULT: https://atlas.web.cern.ch/Atlas/GROUPS/PHYSICS/PAPERS/EXOT-2016-25</a>
GITLAB GROUP	
ID	14349
PROJECTS	
PROJECT	
ID	34774
URL	<a href="https://gitlab.cern.ch/atlas-physics-office/EXOT/ANA-EXOT-2018-01/ANA-EXOT-2018-01-INT1">https://gitlab.cern.ch/atlas-physics-office/EXOT/ANA-EXOT-2018-01/ANA-EXOT-2018-01-INT1</a>
PROJECT	
ID	42146
URL	<a href="https://gitlab.cern.ch/atlas-physics-office/EXOT/ANA-EXOT-2018-01/ANA-EXOT-2018-01-CONF">https://gitlab.cern.ch/atlas-physics-office/EXOT/ANA-EXOT-2018-01/ANA-EXOT-2018-01-CONF</a>
LIKELIHOOD	
INPUT DATA	
PUBLICATIONS	

Copyright 2018 © CERN. Created & Hosted by CERN. Powered by Invenio Software

Figure 19 A CERN Analysis Preservation record that provides the functionality for visualisation (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.

## 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 broadened 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.

Europe PMC

About Tools Developers Help Explore the beta version

Search worldwide, life-sciences literature

HAS\_PUBLISHED VERSION:y

E.g. "breast cancer" HER2 Smith J

[← Back to Results](#)

☐ Identifiers for the 21st century: How to design, provision, and reuse persistent identifiers to maximize utility and impact of life science data (PPR:PPR7049)

This article is a preprint: it has not been peer-reviewed.

Abstract Citations Related Articles Data BioEntities External Links

McMurry J , Juty N , Blomberg N , Burdett A , Conlin T , Conte N , Courtot M , Deck J , Dumontier M , Fellows D , Gonzalez-Beltran A, Gormanns P, Grethe J , Hastings J , Hermjakob H , Heriche J, Ison J, Jimenez R, Jupp S, Kunze J, Laibe C, Le Novère N, Malone JR, Martin M , McEntyre J , Morris C, Muiliu J, Mueller W, Rocca-Serra P, Sansone S, Sanyar M, Snoep J, Stanford NJ, Soiland-Reyes S, Swainston N, Washington N, Williams A, Wimalaratne S , Winfree L, Wolstencroft K, Goble C, Mungall C, Haendel M, Parkinson H

bioRxiv [20 Mar 2017]

Type: Preprint

DOI: [10.1101/117812](https://doi.org/10.1101/117812)

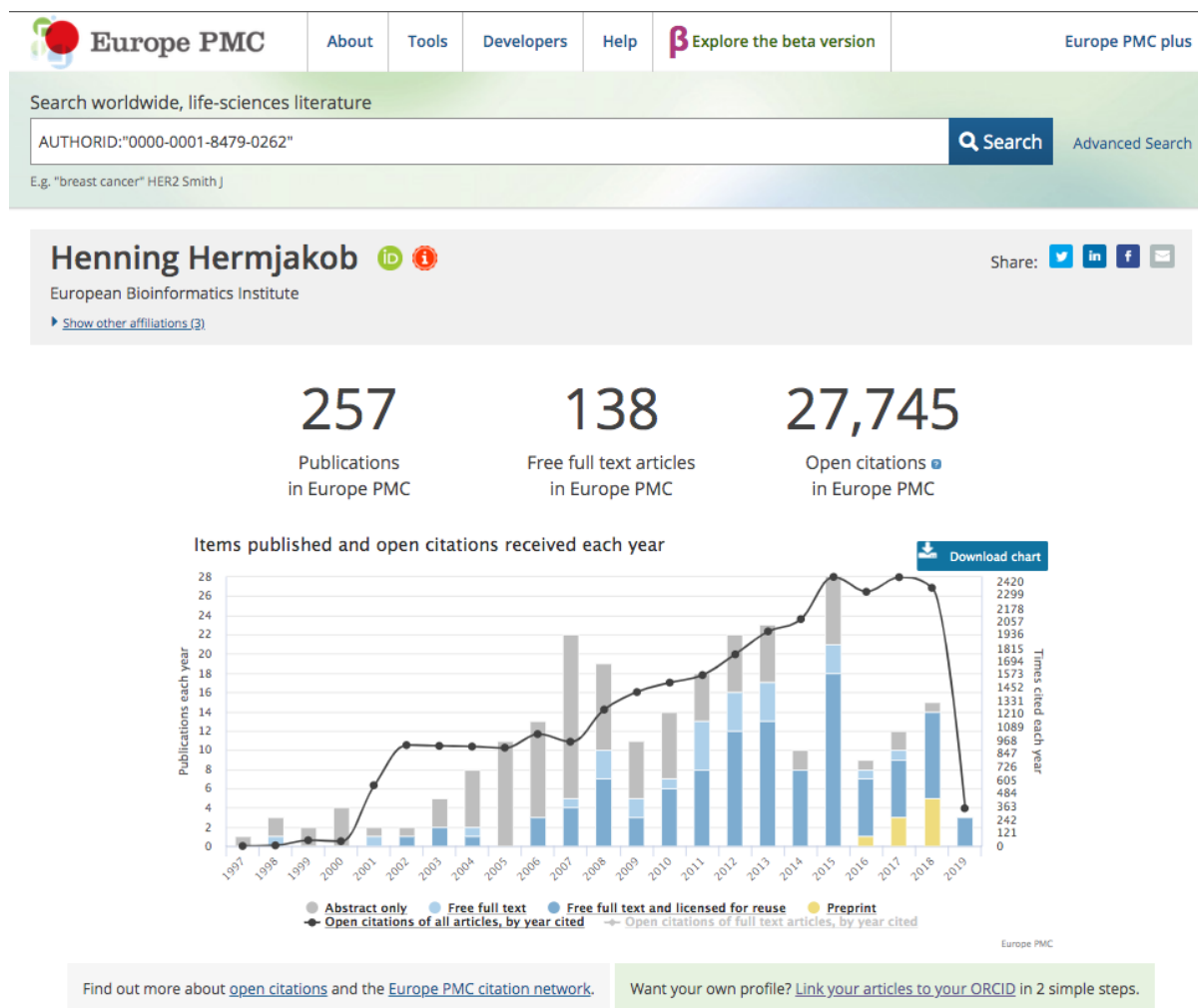
A later version of this preprint was published as "Identifiers for the 21st century: How to design, provision, and reuse persistent identifiers to maximize utility and impact of life science data." PLoS biology. 2017 Jun;15(6):e2001414.

**Abstract**

In many disciplines, data is highly decentralized across thousands of online databases (repositories, registries,

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-coloured 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 itemised 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.*

Europe PMC

About Tools Developers Help [Explore the beta version](#) Europe PMC plus

Search worldwide, life-sciences literature

netsmooth [Search](#) [Advanced Search](#)

E.g. "breast cancer" HER2 Smith J

**Results** [RSS](#) [Save Search](#) [Recent Activity](#) [Export](#)

1 - 5 of 5 results Sort by: Relevance | [Date](#) | [Times Cited](#)

☐ Select results 1 - 5

☐ [netSmooth: Network-smoothing based imputation for single cell RNA-seq](#) [Preprint v3](#)  
 Ronen J, Akalin A  
 F1000Res [10 Jul 2018]  
 Cited: 0 times (PPR:PPR44900)

☐ [netSmooth: Network-smoothing based imputation for single cell RNA-seq](#) [Preprint v2](#)  
 Ronen J, Akalin A  
 F1000Res [24 Jan 2018]  
 Cited: 0 times (PPR:PPR41532)

☐ [netSmooth: Network-smoothing based imputation for single cell RNA-seq](#)  
 Ronen J, Akalin A  
 F1000Res [03 Jan 2018, 7:8]  
 bioc/html/netSmooth.html. Results Overview of the method The intuition behind the **netSmooth** algorithm ... applied **netSmooth**, MAGIC, and scimpute. Figure 4b shows the principal component analysis of **netSmooth**-processed  
 Cited: 0 times (PMID:29511531 PMCID:PMC5814748) [Free full text article](#)

☐ [netSmooth: Network-smoothing based imputation for single cell RNA-seq](#) [Preprint v1](#)  
 Ronen J, Akalin A  
 F1000Res [03 Jan 2018]  
 Cited: 0 times (PPR:PPR41531)

[netSmooth: Network-smoothing based imputation for single cell RNA-seq](#) [Preprint](#)  
 Ronen J, Akalin A  
 bioRxiv [13 Dec 2017]  
 amplifying biases inherent in the data. We present **netSmooth**, a network-diffusion based method that uses priors ...  
 smooth expression values. We demonstrate that **netSmooth** improves clustering results of scRNA-seq experiments  
 Cited: 0 times (PPR:PPR16536)

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 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.

Sign in or create an account

Europe PMC

AboutToolsDevelopersHelp

Explore the beta version

Europe PMC plus

Search worldwide, life-sciences literature

netsmooth

Search

Advanced Search

E.g. "breast cancer" HER2 Smith J

Back to Results

Recent ActivityExportTweet

☐ netSmooth: Network-smoothing based imputation for single cell RNA-seq (PPR:PPR44900)

Formats

AbstractFull Text

This article is a preprint: it has not been peer-reviewed.

AbstractCitationsRelated ArticlesDataBioEntitiesExternal Links

Ronen J, Akalin A

F1000Res [10 Jul 2018]

Type: Preprint

DOI: 10.12688/f1000research.13511.3

Version history

Version 3 [10 Jul 2018]

Version 2 [24 Jan 2018]

Version 1 [3 Jan 2018]

A later version of this preprint was published as "netSmooth: Network-smoothing based imputation for single cell RNA-seq," F1000Research. 2018 ;7:8.

Abstract

No abstract provided.

ORCID*s* linked to this article

Akalin A , 0000-0002-0468-0117 , Max Delbrück Centrum für Molekulare Medizin Berlin Buch

Read Article at publisher's site

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.

## 3.5 Use case: PID Graph for PhD-student activities and outcomes

*FREYA Partner: STFC*

### **PhD research supported by a research funder and large-scale research facilities**

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.

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 scoping exercise, three user stories (#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)

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.

Node	Number of such nodes
EThOS thesis record	578
Person (PhD researcher)	578
STFC institutional repository record	257
Diamond Light Source repository record	228
University of Oxford research archive record	85
Researchfish CRIS record	36
Organization (University)	52
Organization (other)	3
Facility	3
Relationship	Number of such relationships
(Thesis record in repository A)-[:sameThesisAs]->(Thesis record in repository B)	629
(Person)-[:Authored]->(Thesis)	578
(University)-[:AwardedDegreeTo]->(Person)	578
(Person)-[:ExperimentedOn]->(Facility)	461
(Organization)-[:Funds]->(Facility)	2
(Organization)-[:Operates]->(Facility)	2
(Organization A)-[:IsPartOf]-(Organization B)	1

*Table 1 Overview of nodes and relationships derived from the constructed PID graph revolving around student activities*

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?

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 resulted 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.

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. These works will be continued within FREYA WP4 and in part WP3 when it relates to new PID types and provenance. The results of this graph building exercise can be reported through the new Research Graph RDA Interest Group that is going to be established as a follow-up to the Bird-of-Feather Research Graph session that gained a substantial international interest in the 13th RDA plenary meeting in Philadelphia in April 2019.

## 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 through 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 version 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 the 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 to 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.

FREYA’s “BoF” initiative at the 14<sup>th</sup> 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.