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

D4.1 Integration of Mature PID Types

Deliverable type	Report
Dissemination level	Public
Date	30 November 2018

Authors
Artemis Lavasa (CERN, orcid.org/0000-0001-5633-2459)
Sünje Dallmeier-Tiessen (CERN, orcid.org/0000-0002-6137-2348)
Stephanie van de Sandt (CERN, orcid.org/0000-0002-9576-1974)
Ioannis Tsanaktsidis (CERN, orcid.org/0000-0002-1567-3676)
Anna Trzcinska (CERN, orcid.org/0000-0002-5601-2479)
Pamfilos Fokianos (CERN, orcid.org/0000-0003-0618-1722)
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)
Barbara Lemon (BL, orcid.org/0000-0001-6842-0122)
Rachael Kotarski (BL, orcid.org/0000-0001-6843-7960)
Christine Ferguson (EMBL-EBI, orcid.org/0000-0002-9317-6819)
Jo McEntyre (EMBL-EBI, orcid.org/0000-0002-1611-6935)
Simon Lambert (STFC, orcid.org/0000-0001-9570-8121)
Vasily Bunakov (STFC, orcid.org/0000-0003-3467-5690)
Chris Baars (KNAW - DANS, orcid.org/0000-0002-5228-1970)
Maaike de Jong (KNAW - DANS, orcid.org/0000-0003-4803-7411)

Abstract This deliverable reports on the deployment of PID Graph functionality in FREYA's pilot applications. Initial steps for building the PID Graph include advancing the implementation of person-article-data linking, further establishing software citation and publication workflows, and integrating mature PID types into the different disciplinary systems. This first report for Work Package 4 presents considerations and implementations stemming from the first year of work carried out by the pilot applications and sets the scene for the future integration of new and emerging PID types.

Status Submitted to EC 5 December 2018

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

FREYA project summary

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

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

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

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

Disclaimer

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

Executive summary

Persistent Identifiers (PIDs) ensure robust and reliable links to digital objects and between objects. This deliverable demonstrates how community-specific and other existing identifiers can be integrated into disciplinary services. Focusing on PID types that are in a mature state, this report presents some first concrete steps towards building and advancing the individual PID graphs of FREYA's pilot applications. We present how each disciplinary partner created new connections or enhanced already-existing ones during the first year of the project by improving the linking between resources - with a focus on publications, data, and people. During this stage, it was also important to evaluate the degree of advancement with regard to software publishing and the implementation of citation workflows.

The FREYA partners provide specific disciplinary use cases describing current and future integrations of trusted and mature persistent identifiers, as well as community-specific identifiers in established services. The use cases demonstrate differences in terms of community needs and technical requirements amongst the different disciplines and organizations, challenges in PID integration, but also the ability to prototype and implement mature and even emerging PID types despite this disciplinary or organizational heterogeneity.

While the disciplinary use cases show a promising exploitation of PIDs, this is only the starting point. Evaluating the impact of FREYA-driven integrations and supporting the cultural change that is required to fully benefit from these improvements to the research workflows will be challenging, but essential. This deliverable emphasizes the potential of PIDs and how they are crucial to the research workflow in the context of Open Science and the European Open Science Cloud, and presents considerations and possible challenges for the integration of emerging PID types as they further develop in the coming years.

Contents

1 Introduction 6
1.1 Work Package 4 in context 6
1.2 Current landscape and Open Science considerations 7
1.2.1 More data, more complexity, more Open Science? 7
1.2.2 More connections between resources, “better” Open Science? 8
1.2.3 What is the role of PIDs? 8
1.3 User stories in FREYA 9
2 Integrating the PID Graph in FREYA 10
2.1 Introducing the PID Graph 10
2.2 Introducing the pilot applications in FREYA 11
2.3 Developing the pilot applications 12
3 The British Library 13
3.1 Introduction 13
3.2 PID Graphs for EThOS and data.bl.uk 14
3.3 Current implementations 15
3.3.1 data.bl.uk 15
3.3.2 EThOS 16
3.4 Future Implementations 18
3.4.1 data.bl.uk 18
3.4.2 EThOS 19
4 European Organization for Nuclear Research (CERN) 21
4.1 Background and introduction to services 21
4.2 The CERN PID Graph 22
4.3 Established integrations: Building the PID Graph 23
4.3.1 Linking resources 23
4.3.2 Data and software citation workflows in HEP 25
4.3.3 Integration of schema.org 26
4.3.4 ORCID Integration 26
4.3.5 The analysis ID 27
4.4 Integrating new PID types 28
5 Data Archiving and Networked Services (KNAW–DANS) 30
5.1 Introduction to services 30
5.2 The NARCIS PID Graph 30
5.3 Building the PID Graph: Mature PID types 31
5.3.1 ORCID integration 31
5.3.2 Integration of schema.org 32
5.3.3 Unpaywall integration 32

5.3.4 Article-data linking 33
5.3.5 Use cases 33
5.4 Future work: Integrating new PID types 37
6 European Molecular Biology Laboratory (EMBL–EBI) 38
6.1 Background and introduction to services 38
6.2 A literature-centric local PID Graph 39
6.3 Current implementations 41
6.3.1 Adoption of schema.org by Life Science repositories 41
6.3.2 ORCID claiming 41
6.3.3 Data discoverability via BioStudies 42
6.4 Upcoming plans for integrations 42
6.4.1 Grants 42
6.4.2 Research organizations 42
7 PANGAEA 43
7.1 Introduction 43
7.2 The PANGAEA PID Graph 43
7.3 Current PID integrations at PANGAEA 44
7.3.1 Maturity level of PID integration 44
7.3.2 Use cases 44
7.4 Future extensions of the PANGAEA PID Graph 46
8 Science and Technology Facilities Council (STFC) 47
8.1 Introduction 47
8.2 Current STFC graph for PhD theses and related entities 47
8.3 Current PID implementations at STFC 48
8.4 Extensions of STFC graph for PhD theses and related entities 48
9 Conclusions 50
10 Challenges and future work 52
Bibliography 53

1 Introduction

1.1 Work Package 4 in context

The Work Package (WP) of the FREYA project from which this deliverable originates is entitled "Integrating the PID Graph"; since the deliverable itself has the word "integration" in its title, it is worth examining the significance of integration in this context. In particular, it is crucial to specify what it is that is being integrated: the PID Graph is being integrated into disciplinary contexts, that is, the matrices of stakeholders, expectations, practices, workflows (formal and informal), and needs (satisfied and unsatisfied) that are found within particular disciplines and have the degree of commonality that allows reference to a "community" of individuals who identify with the discipline[1]. By "integration of the PID Graph" is meant the clarification, extension, implementation and validation of the idea in the full variety of disciplinary contexts that FREYA offers, with a view to demonstrating that it is not just an attractive hypothetical abstraction, but it has "leverage": the PID Graph is transferable between disciplines and real benefits can be obtained from that transfer[2].

This discipline-focused view of persistent identifiers (PIDs) that FREYA takes is crucial for a few reasons. It is true that the vast majority of uses of persistent identifiers do not relate to a particular domain at all, and indeed have that as a great strength. When the reader of an academic publication clicks on the DOI of one of its citations, knowing that they will instantly be taken to its source (or perhaps be obstructed by a paywall, a less desirable outcome) or clicks on the ORCID iD of one of the authors, with a fair hope of being able to view a concise resumé of that person's career - these are "integrations" of PIDs that provide self-evident benefits for millions of users irrespective of their academic discipline. However, they are not considered as manifestations of the PID Graph that FREYA envisions. At most there might be a couple of hops from one identifier to another, but nothing like the information resource across a network of PIDs that serves as a basis for new services. The resource and the services, if they are to have any depth, must be targeted to a particular discipline; but that does not preclude commonalities arising and being exploited in such cross-disciplinary areas as impact assessment.

To quote the FREYA Description of Activity: "This Work Package is a Service Activity which will develop the PID Graph within specific disciplinary contexts". Its aim is to "build selected full-scale demonstrator systems at TRL7 (through FREYA partners and EOSC) using the PID Graph to illustrate the potential of operational PID services". This deliverable is a first step towards demonstrating how such systems will be actually built through pilot applications. It sets the scene for the disciplinary pilot applications and illustrates how they are being used and will be upgraded thanks to established and new PID services built on discipline-specific PID graphs. Some pressing needs of a wide variety of disciplines are presented, as well as the local PID graphs of each partner. "Integration" here means developing the infrastructure to enable common use of mature PIDs connected through a PID Graph that in the future will be strengthened with new or emerging PID types and new connections.

FREYA's Deliverable 3.1, "Survey of Current PID Services Landscape" (FREYA (2018b)), functioned as a survey on the current state of PID services. The resulting maturity matrix demonstrates that data publishing services and article–data linking are technically in a mature state. At the same time, PID services for software are emerging and gaining a growing importance. The report also demonstrated that even mature PID services, e.g. for research data, need further development as existing solutions are not yet pervasive and need further development and community integration.

[1] The words "field", "domain", "discipline" and "community" are used synonymously here.

[2] In fact, this Work Package also calls for integration with the European Open Science Cloud (EOSC), but that sense of integration may be rather different and is not the direct concern of this first deliverable.

The following sections in this introduction give an overview of the current PID and Open Science landscape, both from a general and a disciplinary point of view. Following that, the FREYA concepts “PID Graph”, “pilot applications” and “user stories” are explained to better illustrate what Work Package 4 builds on and the way this is accomplished. Chapters 3 to 8 present the work that has been carried out so far by each disciplinary partner, i.e. the British Library, CERN, DANS, EBI, PANGAEA and STFC.

The emphasis is initially on mature PID types (as identified in Deliverable 3.1): articles, datasets and people, though also included are software (publishing and citation workflows), as well as mature and trusted community-relevant identifiers. However, the discussion of each of the pilot applications distinguishes between current and future implementations, and the latter may extend to include PID types currently considered “immature”. Subsequent work in FREYA will concentrate in particular on provenance in the context of disciplinary systems, as well as implementation of advanced PID Graph functionality to exploit the richness of connections between diverse PID types. The conclusions at the end of this deliverable highlight challenges and considerations regarding the next steps for the pilot applications as part of this Work Package.

1.2 Current landscape and Open Science considerations

1.2.1 More data, more complexity, more Open Science?

One factor that unites almost all disciplines of academic study is the growing importance of data. We live in times of highly data-intensive science discoveries where hundreds of petabytes of data is produced on a single day (Hey et al. (2009)). Not only has the storage size and amount of research objects increased (Maass et al. (2017)), but the research landscape is more complex and granular than ever before (Hey et al. (2009)), with many different research entities combining to contribute to the overall research picture (FREYA (2018b)).

Looking at the current landscape of data production, the trend towards data-driven research (Anderson (2008)) and increasing data volumes, variety and velocity (“Big Data”) is visible across disciplines (Maass et al. (2017)). For example, particle collisions measured by the Large Hadron Collider (LHC) at CERN produce about one petabyte of collision data per second. In June 2017, CERN’s data center reached the 200 petabyte milestone of data permanently archived in the tape library[3]. Astronomers are now confronted with getting their answers from queries to databases containing petabytes of data resulting from space observations (Murray (2017)).

Similar trends can be observed in the Humanities and Social Sciences; longitudinal and panel data in the Social Sciences are intricate and large in size. Complex repeated measurements taken from a cross-section of subjects in Biology and the Social Sciences enforce new methods to handle them (Frees (2004)). Entire new research fields appeared out of this shift in research and social scientists could now be considered “data scientists” (Foster et al. (2016)). Furthermore, numerous research projects under the banner of the Digital Humanities (DH) combine traditional research in the Humanities with computing and digital advantages (Drucker (2013)). The digitization enables the emergence of projects like the Digital Pantheon[4], which allows studying and annotating 3D models and isometric projections of ancient architecture. Manuscripts can now be digitally encoded and annotated in huge corpora. Finally, the field of Corpus Linguistics evolved as well, as rigorous annotating in a digital environment can lead to a greater linguistic understanding (Wallis (2007)). The increasing amount of data in combination with new computational methods enable data-driven discoveries that were not possible without computational power (Dzogang et al. (2016)).

[3] 200 petabytes in the CERN Data Centre: https://home.cern/about/updates/2017/07/cern-data-centre-passes-200-petabyte-milestone

[4] Digital Pantheon: http://repository.edition-topoi.org/collection/BDPP

As demonstrated, the basis for this evolving research landscape is data. The existence of high-quality research data is crucial for conducting research in the current landscape. With the increasing urge for high-quality research data comes the compelling necessity for freely-accessible and findable data. The Open Science movement encourages scientists to publish and communicate their knowledge without barriers (Molloy (2011)). Open Science practices are being adopted more broadly across disciplines and stakeholder types and that becomes clear when examining the current landscape of data management in research organizations or any institution that produces data (Miguel et al. (2014); Nosek et al. (2015); Sitek & Bertelmann (2014); Pampel & Dallmeier-Tiessen (2014)). This shift towards Open Science practices reduces the cost and difficulty of data handling and may advance the state of research and innovation (Borgman (2012)).

1.2.2 More connections between resources, “better” Open Science?

Many resources are publicly-available as a result of the adoption of Open Science practices and in many cases can be used by others. Taking a closer look at publishing approaches where reusability is an essential element, we can observe that their advantage is that they provide context. In order to understand the whole picture of the research process, the context surrounding any given object needs to be available as well. The availability of just the data is not enough, as data is often difficult to interpret once removed from its initial context. One of the key benefits of Open Science is the reusability of publicly-available research objects, which allows third parties to ask new questions by using data created by others. The FAIR data principles (FORCE 11 (2016)) suggest that data has to be associated with their provenance in order to be reusable. Therefore, data has to be linked to all associated objects that reveal the whole story behind any given object to give users all the information they need (context) and to enable reusability.

Such links already exist to some extent. Links between resources are created every day. A researcher may work on a research project funded by a specific funding agency, producing data and code to solve a research problem, and several papers can be published as a result. So, in this example, all resources (i.e. the data, the software, the publication, the researcher’s organization and the funder) are connected through the researcher. The whole research process could be described as a network of events and resources where the key knots (links) are connecting every other dot (research resources like datasets) in the network, revealing the provenance and context of a research project or research result. These connections are a fundamental contribution to modern research and facilitate findability and reusability.

Reusability has been the topic of discussion in many disciplines (Open Science Collaboration (2015); Bakeer (2016)). Reusability - or reproducibility depending on the definition - requires, amongst other information, at least the underlying data and software, which are often publicly available as well. The problem here is the missing link between all the resources that belong together. Even in cases where all the related resources are findable and accessible, it is not uncommon that they are not linked to one another.

1.2.3 What is the role of PIDs?

The aforementioned connections are established by using PIDs to persistently and uniquely identify digital objects or digital representations of organizations, people, etc. Crossref[5] and DataCite[6], for example, are Digital Object Identifier (DOI)[7] registration agencies for written works, data and other resources, whose purpose is to enhance the linking of research resources via PIDs. ORCID[8] offers the opportunity to uniquely identify researchers by PIDs, which solves not only name ambiguity but also contributes towards better connectivity between resources as researchers can link publications to their user profiles. PIDs of related resources are linked together by including any relevant PIDs in the metadata of the resource and/or in the metadata of the identifier itself. Regarding the latter, an example would be adding identifiers of related

[5] Crossref: https://www.crossref.org/

[6] DataCite: https://www.datacite.org/

[7] DOI: https://www.doi.org/

[8] ORCID: https://orcid.org/

resources to the “relatedIdentifier” property when registering DataCite DOIs. Several more services contribute to make this vision a reality.

In summary, PIDs bring structure and trust into a complex research landscape. This does not only concern the FREYA pilot applications, but any Open Science service, from repositories to aggregators. Particularly noteworthy is the European Open Science Cloud (EOSC) (European Commission (2017)) which also relies on PIDs as the backbone of their services.

1.3 User stories in FREYA

In the context of the massive growth in data and connections that has just been outlined, it is necessary for FREYA to adopt a consistent approach. This is done through the use of user stories arising from the needs of the various subject disciplines represented in the project, which is part of the work carried out in Work Package 3. A user story is a technique of requirements acquisition for software or product development, focusing on the perspectives of end users of the system and intended to capture their desires using a simple template, such as:

As a <role>, I want <capability> so that <benefit>.

In this way, the FREYA partners can present and analyze important needs from their respective domains with a consistent starting point, allowing the diversity of needs to be captured and at the same time allowing the possibility of common requirements across disciplines.

Work Package 3 continues to build on this approach as part of the upcoming Deliverable 3.2; through collecting, analyzing and prioritizing use cases, we will be able to understand the requirements for the implementation of new PID services.

2 Integrating the PID Graph in FREYA

2.1 Introducing the PID Graph

The PID Graph is one of the driving visions of the FREYA project, along with the PID Forum and the PID Commons. The PID Forum is a stakeholder community, whose members collectively oversee the development and deployment of new PID types; the PID Commons is concerned with the sustainability of the PID infrastructure beyond the lifetime of the project, and comprises the roles, responsibilities and structures for good self-governance based on consensual decision-making. The PID Graph is the unifying vision of the current and future PID landscape brought together to create and exploit networks of research-related entities associated with PIDs.

The PID Graph has already been mentioned in previous chapters of this deliverable, with hints at some of its attributes: it is to be a basis for new services in multiple domains, yet in itself it is domain-independent; it is also capable of encompassing many types of entities with PIDs. The PID Graph is a complex concept and not susceptible to a simple definition. Instead, it can be viewed from several complementary and interlinked perspectives, each of which illuminates an aspect of the Graph that will be important for particular areas of the work of FREYA and for the PID infrastructure in general. To better understand what the PID Graph means in the context of FREYA, it would perhaps be helpful to divide it into three elements, which could be labelled *vision*, *content provision* and *supporting infrastructure*.

The vision means establishing and exploiting the connections between entities in the research domain (through their PIDs) as a basis for applications responding to needs and deriving value for users. FREYA and the PID Graph focus on connections: the emphasis is not primarily on PIDs for the purpose of identification of (digital) objects, though that role is certainly a necessary one. This vision leads to the next perspective (content provision), since the needs and the value arise in particular disciplinary contexts and for particular purposes. That implies the existence of individual PID graphs over which the applications operate, constructed and maintained for the benefit of the individual disciplinary communities. Of course the actual elements of the graphs (the entities with PIDs) are not specific to the applications but have some generality: publications, datasets, people, software, instruments and so on.

Because of the universality of such entities across most fields of science, it is tempting to envisage a universal PID Graph, but that would be practically infeasible and in fact there is no real need. The PID Graph is not to be thought of as a single entity, but as the federation of the local graphs built on the supporting infrastructure for particular applications. These graphs taken together make up the federated PID Graph. According to the vision of FREYA, there will be clear uniformity across these local graphs, and they could in principle be linked together into larger ensembles.

This leads on to the third perspective, the supporting infrastructure of the PID Graph: what is needed to build those distributed local graphs. From FREYA's point of view, the mechanisms of construction and use of the local graphs are of as much importance as the graphs themselves. Best practices, standards and APIs will enable local graphs to be built efficiently and with rich content and therefore, through this common foundation, form part of the overall PID Graph.

These three perspectives all make up the overall PID Graph. The working out of the implications of these three perspectives is precisely the job of FREYA, answering questions such as:

- How are the local graphs built, and who develops the applications that operate over them?
- What are the components of the supporting infrastructure? Who defines them, who has to implement them?
- What do PID service providers need to do to contribute to the supporting infrastructure?

An incidental issue relevant to this deliverable is the interpretation of the word “graph” in an application or domain context. It is worth stating for clarity that there are two steps that are both required in the development of a graph, and to which the word “graph” could validly be applied. One is the design of an abstract graph linking types of entities with their PIDs. The purpose is to map out what PIDs and what connections are needed to enable the giving of value responding to community needs. This may be a depiction in any convenient form, not necessarily a machine-usable form, and is what is presented in this deliverable for the pilot applications. A possible next step could be an actual implementation of that abstract graph, populated with content and existing in some machine-usable form ready to be operated over by an application.

Taking all of the above into account, the PID Graph can be considered a network of connections between PIDs available through a set of federated RESTful APIs. The present deliverable takes an application-centric view, driven by needs of particular domains and user groups, to design the graphs that will together form part of the overall PID Graph - to integrate the PID Graph into disciplinary contexts.

2.2 Introducing the pilot applications in FREYA

The FREYA partners represent a range of academic disciplines:

- The British Library: Humanities and Social Sciences
- CERN: High-Energy Physics
- KNAW - DANS: Social Sciences
- EMBL - EBI: Life Sciences
- PANGAEA: Earth and Environmental Sciences
- STFC: Facilities-based Science

These partners are providing pilot applications for PIDs within FREYA. It is worth clarifying what is meant by the term “pilot application” in this context. On the one hand it refers to a general area of need within a particular discipline: for example, the assessment of impact of the use of research facilities such as those operated by STFC or enhancing reproducibility through provenance of datasets. More specifically, a pilot application in FREYA is a concrete development, built on the PID infrastructure and exploiting a particular PID graph, that will meet the needs identified within a discipline. Identifying user needs allows the transition from the general to the specific: from the overall need within a disciplinary context to a particular PID graph whose implementation will respond to that need.

The pilot applications will show how user-motivated systems can be integrated with the PID Graph and the variety of ways to develop trusted PID Graph services, providing best practice examples and documentation. Furthermore, the pilot applications have an important role in the dissemination of FREYA’s work. Communities can identify with needs, see benefits, and adopt best practices in their own applications.

The initial goal is to focus on mature PID services. In this deliverable, the individual approaches of the disciplinary pilot applications within FREYA are introduced and the various interests and stages of implementation of different PID types are surfaced. While certain PID types may be considered “emerging” or “immature” in general, individual fields or even institutions have achieved a mature use of PID systems beyond those identified in Deliverable 3.1, namely PIDs for publications, data and authors (researchers). The pilot applications will help demonstrate the maturity that has been achieved by the individual partners and the potential impact through wider implementation by similar institutions. Further updates on the work presented here will be given in future Work Package 4 deliverables.

2.3 Developing the pilot applications

In general, the development of all FREYA's pilot applications as part of Work Package 4 will pass through a number of stages. It is important to note that FREYA Work Packages are interlinked, in that the work carried out by other Work Packages feeds into the tasks required for the completion of the individual stages indicated in the table below. Furthermore, the engagement Work Package (WP5), even though not explicitly shown in the table, plays a key role in all these activities as it continuously gathers feedback from the wider community (PID Forum) which is used to inform FREYA's work.

Hence, the following table (Table 1) illustrates the stages of development for the pilot applications, where for each stage the corresponding deliverable is specified, as well as any connections to other Work Packages:

Deliverable	Stage of development	Connections to other WPs
D4.1	**Constructing disciplinary PID Graphs**	**It includes considerations regarding PID maturity levels based on WP3's work (D3.1)**
	Building up with mature PID types	
D4.2, D4.3 & D4.7	Enhancing PID Graph functionality	It will build on future work by WP2
D4.5	Scaling up PID Graph integrations (EOSC)	It will feed into WP6
D4.4 & D4.6	Augmenting with emerging/new PID types	It will build on future work by WP3

Table 1: Stages of development for the disciplinary pilot applications in Work Package 4 (stages corresponding to the present deliverable in bold letters).

The present deliverable, D4.1, with its focus on presenting and building on the disciplinary graphs for mature PID types, covers the first two of these stages and future deliverables will cover the rest.

3 The British Library

3.1 Introduction

As the national library for the United Kingdom, the British Library[9] holds vast collections including a continually expanding range of digital materials, such as digitized manuscripts, audio files, newspaper archives, journal articles, research reports, policy documents, web archives, research data and academic theses. These materials cover all subject areas and disciplines, but for the purposes of the FREYA project, the British Library opted to focus on piloting integrations in platforms that provide resources for Arts and Humanities, in particular its research datasets and doctoral theses.

The first platform is data.bl.uk, designed to make selected British Library datasets available for research and creative purposes. The collection includes over one hundred datasets ranging from catalogue records of 19th century Indian books, digitized Hebrew manuscripts, medieval maps, and 17th century theater playbills through to a contemporary UK web archive.

The second platform is EThOS[10], the British Library's long-running database of doctoral theses awarded by UK universities. Over 500.000 titles from more than 120 universities are listed in EThOS.

The British Library is currently leading development and testing of a shared institutional repository with the Tate, British Museum, Museum of London Archaeology and National Museums Scotland. Among the specifications for the repository, being developed by Ubiquity Press, is a requirement for support of multiple PID types within the metadata schema and for functionality to mint DOIs for repository content. Pending the success of the pilot, the British Library intends to migrate content from both EThOS and data.bl.uk into this more permanent repository.

The British Library does not host its own persistent identifier platform, but currently registers DOIs via DataCite.

[9] The British Library: https://www.bl.uk/
[10] EThOS: https://ethos.bl.uk

3.2 PID Graphs for EThOS and data.bl.uk

The graphs below (Figures 1 & 2) illustrate PID types currently recognized and incorporated into data.bl.uk and EThOS.

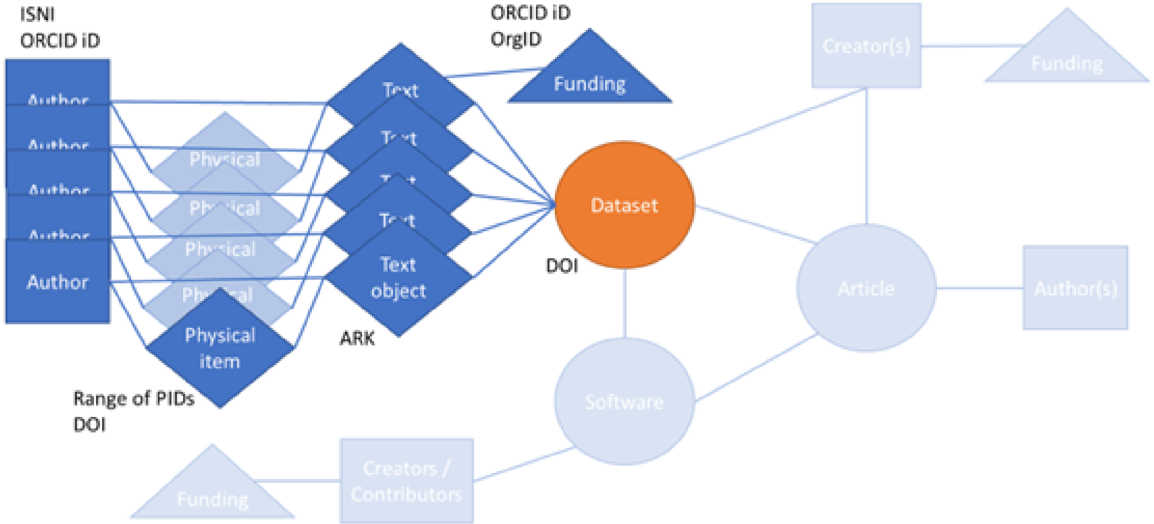


Figure 1: PID Graph representing metadata in data.bl.uk currently connected by persistent identifiers. Fields faded out have the potential to be connected but are not as yet. Note that the provenance of each dataset comes from the items that are collected to create it, e.g. digitized texts. Each of these has its own author and also relates to a physical item.

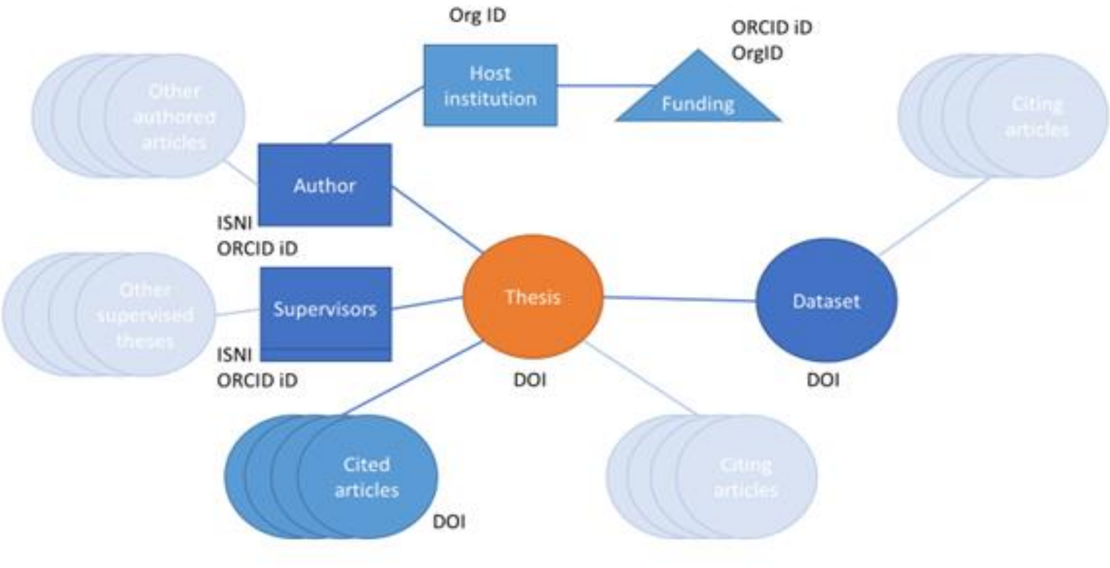
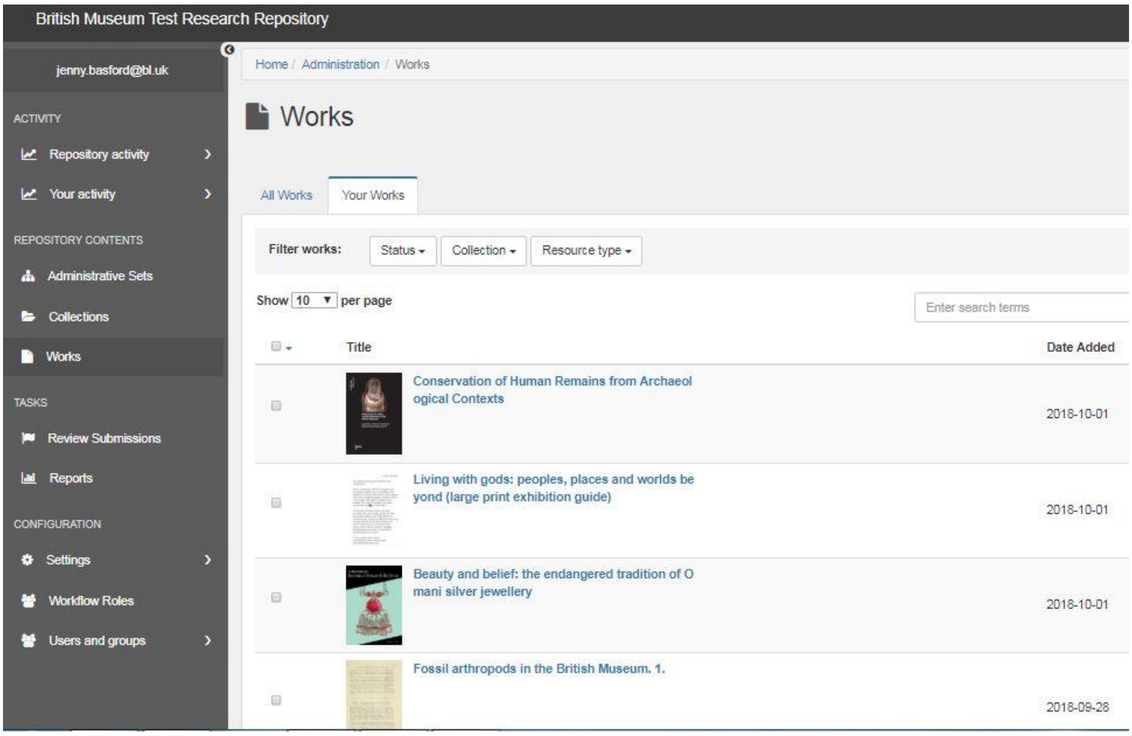


Figure 2: PID Graph representing metadata currently available in EThOS. Darker cells represent metadata directly received by EThOS, with lighter cells showing metadata that can be extracted from full-text theses. Faded cells represent metadata only available via the global PID Graph, i.e. beyond the British Library.

3.3 Current implementations

3.3.1 data.bl.uk

The British Library began a 15-month pilot for a shared repository in June 2018 with four partner institutions. The aim was to assess the cost and share the development expertise required to build a specialist repository for cultural collections, and to test the design against a range of collection types and formats. Each partner institution has identified priority content with which they would like to test integration, including book publication series, data, journal articles and reports (Figure 3).



Title	Date Added
Conservation of Human Remains from Archaeological Contexts	2018-10-01
Living with gods: peoples, places and worlds beyond (large print exhibition guide)	2018-10-01
Beauty and belief: the endangered tradition of Omani silver jewellery	2018-10-01
Fossil arthropods in the British Museum. 1.	2018-09-28

Figure 3: Screenshot of the dashboard for the shared repository (in pilot).

The repository is being developed by Ubiquity Press using the Samvera Hyku repository application. A minimal viable product is due for completion in November 2018.

Ahead of the migration of content, all partners without their own repository (i.e. all but National Museums Scotland) have been provided with a template that stipulates the metadata compulsory for assignment of a DOI. Related identifiers can be added in a free text field. Separate fields are provided for ISSNs and ISBNs. These fields will all be available in the minimum viable product.

Datasets within data.bl.uk have been prepared with the addition of ISNIs and ORCID iDs for creators, contributors, editors and organizations.

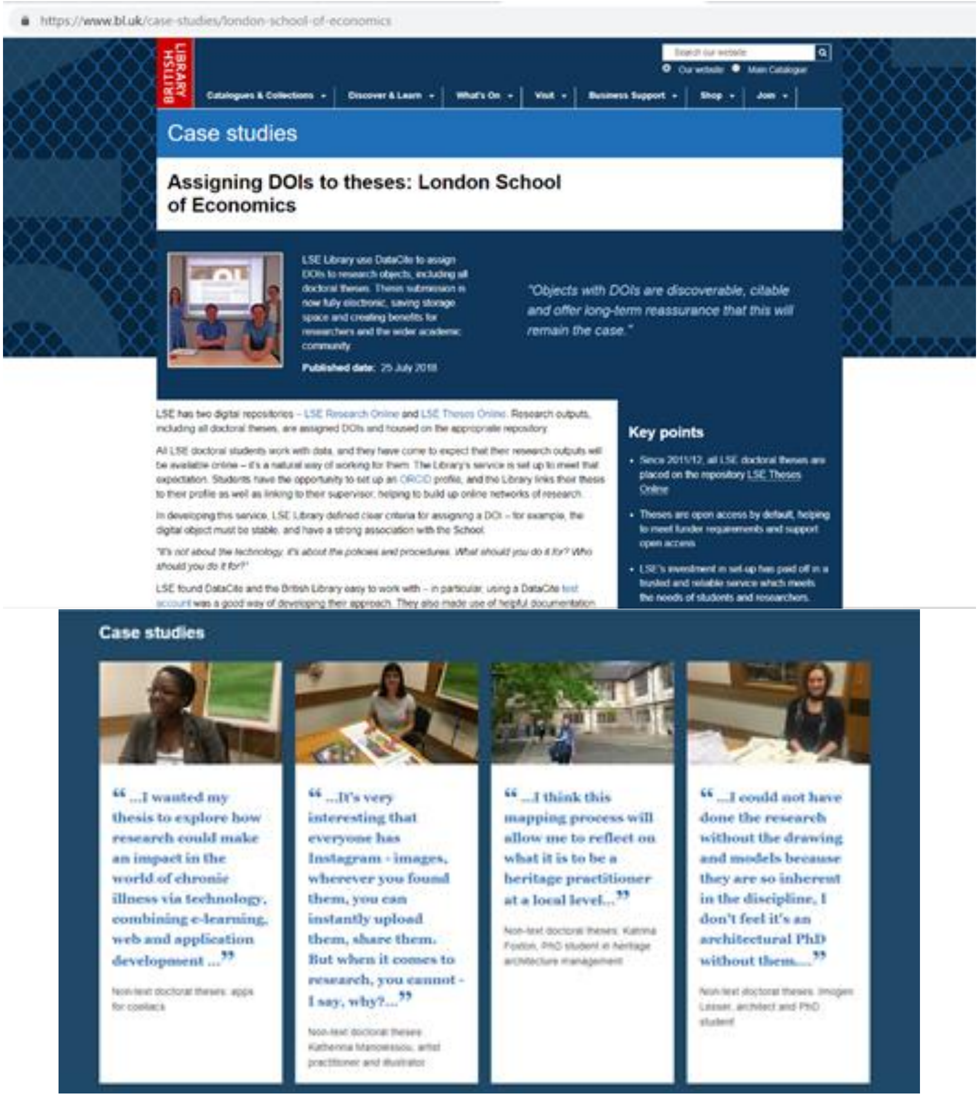
One of the most useful attributes of the repository will be its in-built functionality for minting DOIs using the DataCite API. This is due for testing in November 2018. The DataCite integration means that items within the repository can be assigned a DOI in situ and linked more effectively with ORCID iDs and ISNIs in their own metadata.

The repository is not yet equipped for storage of software, nor for software to be previewed within the browser function, but this will be revisited in later iterations. Currently the search function within the repository is separated per institution but a shared layer for comprehensive search across all institutional holdings is being developed.

3.3.2 EThOS

EThOS is the UK's national thesis service and provides an aggregated record of all doctoral theses awarded by UK universities. Of the more than 500.000 titles listed in this database, approximately half (270.000) are available for free digital download, either from the EThOS database itself or via links to institutional repositories. Those unavailable for download are primarily older theses held only in print form, with database records dating back as far as 1780. EThOS offers a digitization-on-demand service for theses in this category.

The EThOS platform is ageing and is due for migration into the shared repository in 2020-2021. A test migration of metadata is scheduled for early 2019, with a trial migration of EThOS content to follow in the second half of 2019. In preparation, the metadata schema behind EThOS (UKETD_DC[11]) has been expanded to accommodate a wider range of persistent identifiers. Currently these are ISNI, ORCID iD and DOI. A new field for subject classification is under development, with further plans to accommodate organization IDs and research facility IDs.



Case studies

Assigning DOIs to theses: London School of Economics

LSE Library use DataCite to assign DOIs to research objects, including all doctoral theses. Theses submission is now fully electronic, saving storage space and creating benefits for researchers and the wider academic community.

Published date: 25 July 2018

"Objects with DOIs are discoverable, citable and offer long-term reassurance that this will remain the case."

LSE has two digital repositories – LSE Research Online and LSE Theses Online. Research outputs, including all doctoral theses, are assigned DOIs and housed on the appropriate repository.

All LSE doctoral students work with data, and they have come to expect that their research outputs will be available online – it's a natural way of working for them. The Library's service is set up to meet that expectation. Students have the opportunity to set up an ORCID profile, and the Library links their thesis to their profile as well as linking to their supervisor, helping to build up online networks of research.

In developing this service, LSE Library defined clear criteria for assigning a DOI – for example, the digital object must be stable, and have a strong association with the School.

"It's not about the technology, it's about the policies and procedures. What should you do it for? Who should you do it for?"

LSE found DataCite and the British Library easy to work with – in particular, using a DataCite test account was a good way of developing their approach. They also made use of helpful documentation

Key points

- Since 2011/12, all LSE doctoral theses are placed on the repository LSE Theses Online
- Theses are open access by default, helping to meet funder requirements and support open access
- LSE's investment in set-up has paid off in a trusted and reliable service which meets the needs of students and researchers.

Case studies

"...I wanted my thesis to explore how research could make an impact in the world of chronic illness via technology, combining e-learning, web and application development ..."

Non-text doctoral theses: apps for coeliacs

"...It's very interesting that everyone has Instagram - images, wherever you found them, you can instantly upload them, share them. But when it comes to research, you cannot - I say, why?..."

Non-text doctoral theses: Katherina Manoessou, artist practitioner and illustrator

"...I think this mapping process will allow me to reflect on what it is to be a heritage practitioner at a local level..."

Non-text doctoral theses: Katrina Foxton, PhD student in heritage architecture management

"...I could not have done the research without the drawing and models because they are so inherent in the discipline, I don't feel it's an architectural PhD without them...."

Non-text doctoral theses: Imogen Lesser, architect and PhD student

Figure 4: Two screenshots from the British Library website including the EThOS toolkit and case studies.

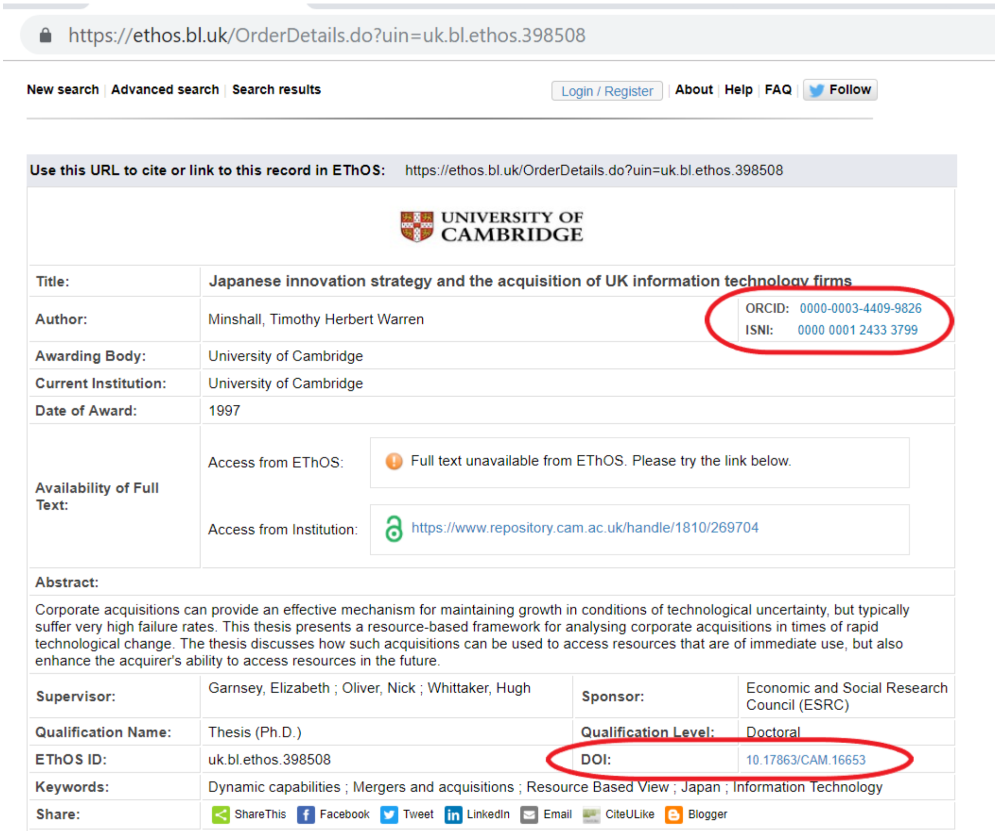
[11] EThOS metadata schema: http://ethostoolkit.cranfield.ac.uk/tiki-index.php?page=The+EThOS+UKETD_DC+application+profile

In order to increase the number of theses with an assigned DOI, the British Library conducted an advocacy and outreach program including provision of resources for university repositories and for thesis authors themselves. Part of that was a series of case studies setting out how to assign a DOI to a thesis and explaining the benefits of persistent identifiers for researchers in all disciplines (Figure 4).

As of September 2018, the EThOS database included:

- 5019 DOIs (from 22 institutions)
- 1149 ORCID iDs (from 43 institutions)
- 386705 ISNIs

A total of 121 records in EThOS contained all three PID types. In these cases, the DOI applies only to the electronic copy of the thesis, while both the ISNI and ORCID iD are attached to the thesis author (Figure 5).



https://ethos.bl.uk/OrderDetails.do?uin=uk.bl.ethos.398508

New search | Advanced search | Search results
Login / Register | About | Help | FAQ | Follow

Use this URL to cite or link to this record in EThOS: https://ethos.bl.uk/OrderDetails.do?uin=uk.bl.ethos.398508

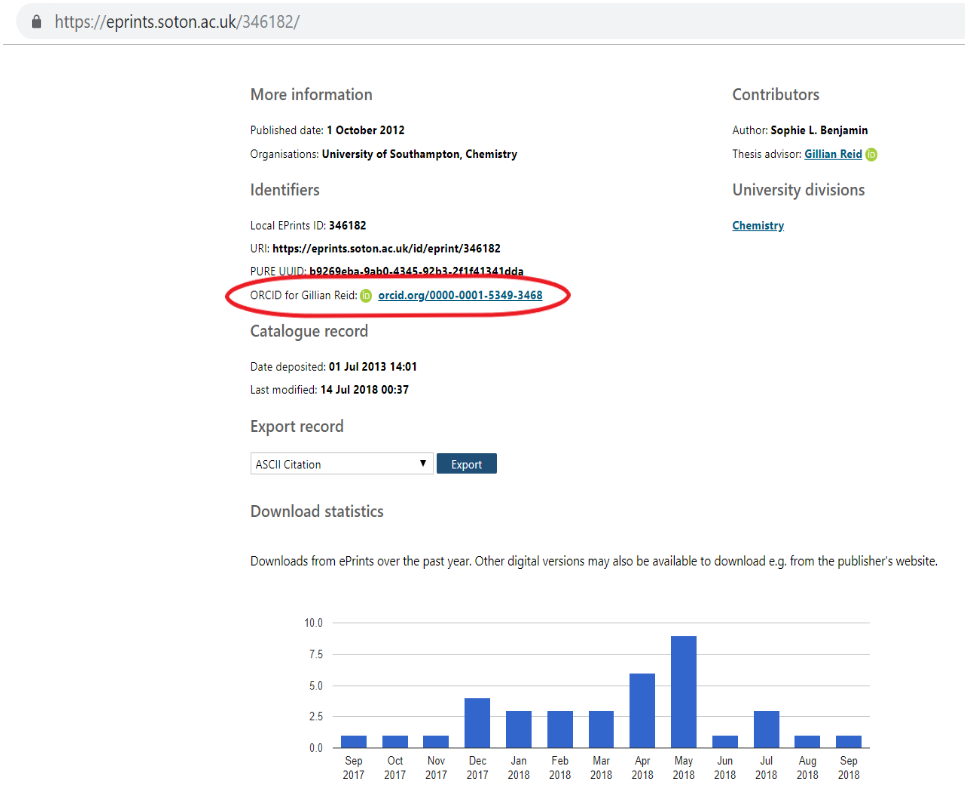
UNIVERSITY OF CAMBRIDGE

Title:	Japanese innovation strategy and the acquisition of UK information technology firms		
Author:	Minshall, Timothy Herbert Warren	ORCID: 0000-0003-4409-9826 ISNI: 0000 0001 2433 3799	
Awarding Body:	University of Cambridge		
Current Institution:	University of Cambridge		
Date of Award:	1997		
Availability of Full Text:	Access from EThOS:	Full text unavailable from EThOS. Please try the link below.	
	Access from Institution:	https://www.repository.cam.ac.uk/handle/1810/269704	
Abstract:			
Corporate acquisitions can provide an effective mechanism for maintaining growth in conditions of technological uncertainty, but typically suffer very high failure rates. This thesis presents a resource-based framework for analysing corporate acquisitions in times of rapid technological change. The thesis discusses how such acquisitions can be used to access resources that are of immediate use, but also enhance the acquirer's ability to access resources in the future.			
Supervisor:	Garnsey, Elizabeth ; Oliver, Nick ; Whittaker, Hugh	Sponsor:	Economic and Social Research Council (ESRC)
Qualification Name:	Thesis (Ph.D.)	Qualification Level:	Doctoral
EThOS ID:	uk.bl.ethos.398508	DOI:	10.17863/CAM.16653
Keywords:	Dynamic capabilities ; Mergers and acquisitions ; Resource Based View ; Japan ; Information Technology		
Share:	ShareThis Facebook Tweet LinkedIn Email CiteULike Blogger		

Figure 5: Screenshot from an EThOS catalogue record including ORCID iD, ISNI and DOI.

To date, further application of ORCID iDs to supervisors and associated researchers has been inhibited by the arrangement of metadata within the UKETD_DC schema. Author and supervisor names are harvested separately from their identifiers, which makes re-pairing problematic, if not all authors and supervisors have an ORCID iD and therefore don't "line up". The schema does not currently allow for the explicit linking of multiple person names with multiple person identifiers.

Already, ORCID iDs and other PID types are becoming more widely adopted by universities in the UK and these identifiers will be included in thesis data received by EThOS. The University of Southampton, for example, has begun including ORCID iDs for supervisors in their own repository records (Figure 6).



https://eprints.soton.ac.uk/346182/

More information

Published date: **1 October 2012**

Organisations: **University of Southampton, Chemistry**

Identifiers

Local EPrints ID: **346182**

URI: **https://eprints.soton.ac.uk/id/eprint/346182**

PURE UUID: **b9269eba-9ab0-4345-92b3-2f1f41341dda**

ORCID for Gillian Reid: orcid.org/0000-0001-5349-3468

Catalogue record

Date deposited: **01 Jul 2013 14:01**

Last modified: **14 Jul 2018 00:37**

Export record

Download statistics

Downloads from ePrints over the past year. Other digital versions may also be available to download e.g. from the publisher's website.

Contributors

Author: **Sophie L. Benjamin**

Thesis advisor: Gillian Reid

University divisions

Chemistry

Figure 6: From the University of Southampton Institutional Repository. Record for Sophie L. Benjamin (2012), Synthesis and coordination chemistry of hybrid polydentate and halide-substituted stibines and bismuthines.

It is anticipated that the question of attaching identifiers for multiple researchers within the same record will be resolved within the repository metadata schema, following a full migration of EThOS content. Links between authors, supervisors and associated researchers will allow for the development of researcher "family trees", in turn providing a far richer picture of research networks and the impact of doctoral research in the UK.

3.4 Future Implementations

3.4.1 data.bl.uk

Given the size of the datasets within data.bl.uk, one potential development of interest to the British Library is producing and linking to data mining tools for researchers. This is particularly important for researchers in the Digital Humanities, given that methods and tools for data analysis at scale tend to be the domain of social scientists, data scientists and information systems experts.

As part of a series of student placements, the British Library has investigated the possibility of developing Jupyter Notebooks that can be connected to specific datasets (such as a package of digitized 19th century novels) to assist researchers with tasks such as applying an API, navigating compressed files, understanding basic Python language, using algorithms or applying text analysis techniques that can generate word frequencies, basic charts and topic modelling.

The difficulty is partly in the linking of these Jupyter Notebooks to a dataset – at this stage only possible using the free text "relatedIdentifier" field – and partly in maintaining currency of the Notebooks. There are unresolved questions such as to whether a Notebook could itself be assigned a DOI, how to ensure it is hosted reliably, and how to review it over time to ensure its content remains technically up to date and relevant to the dataset in question.

Supporting discovery of British Library datasets is also an important role of the repository. One of the key elements to enable discovery will be eventual support for embedded schema.org metadata within repository landing pages. There are two possible approaches to this work, using the DataCite Search API to format and embed schema.org-compliant metadata into the landing pages, or formatting the repository metadata locally. The best approach will be investigated as the repository develops further.

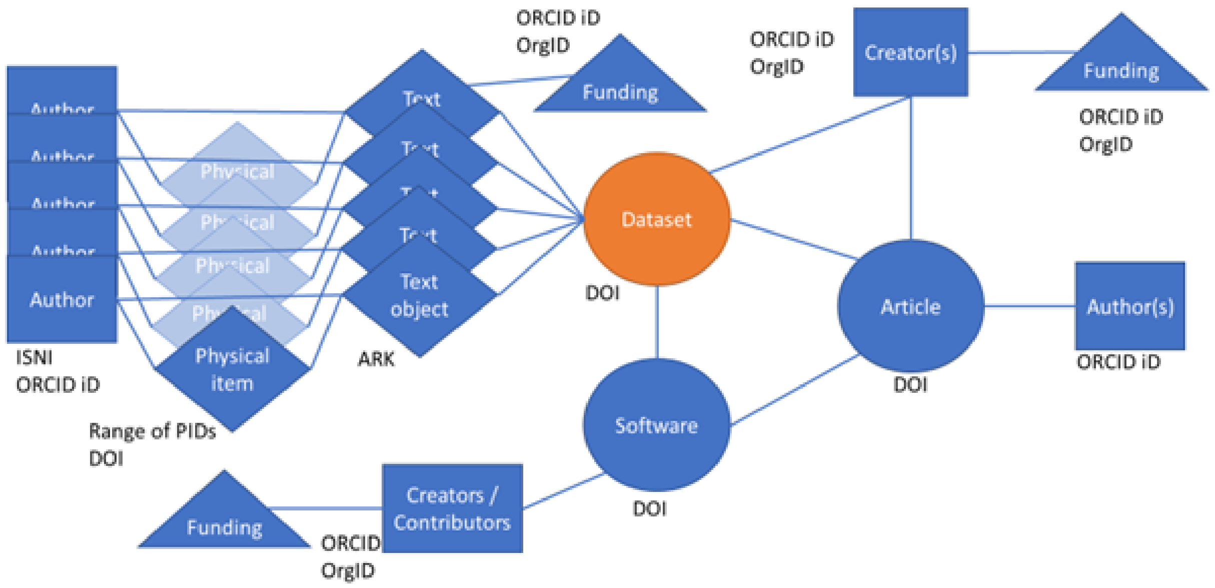


Figure 7: PID Graph representing metadata in data.bl.uk with the potential to be connected by persistent identifiers.

3.4.2 EThOS

EThOS is a comprehensive database with the capacity to provide direct access to original research, but also to information about research using the metadata behind doctoral thesis records. From this metadata we can learn a lot about the nature and extent of research networks, reuse of research data, dissemination, citation, influence and impact.

The British Library is currently engaged in a collaborative project with King's College that will use machine learning to analyze metadata extracted from the first ten pages of every digitized thesis in the EThOS database, including names of funding bodies, supervisors and associated researchers.

The MODS (Mapping Knowledge with Data Science) project is based upon the premise that the PhD constitutes a formative period in the development of the individual researcher, positioning them within an academic "lineage" that continues with their own students. Analysis of bibliographic data at scale is expected to reveal patterns in the flow of ideas and the ways in which they are generated between supervisors, students, departments and institutions. This can be extended to analyze the role of institutions in promoting and sustaining innovation, and the contributions to knowledge transfer made by individuals who move between institutions.

Given the scale of analysis required, this project will focus on two to three academic domains (e.g. Geography, Artificial Intelligence, History). There is a great deal of potential for future projects of this nature given that data extraction will be simplified by the consistent inclusion of multiple persistent identifier types in thesis metadata.

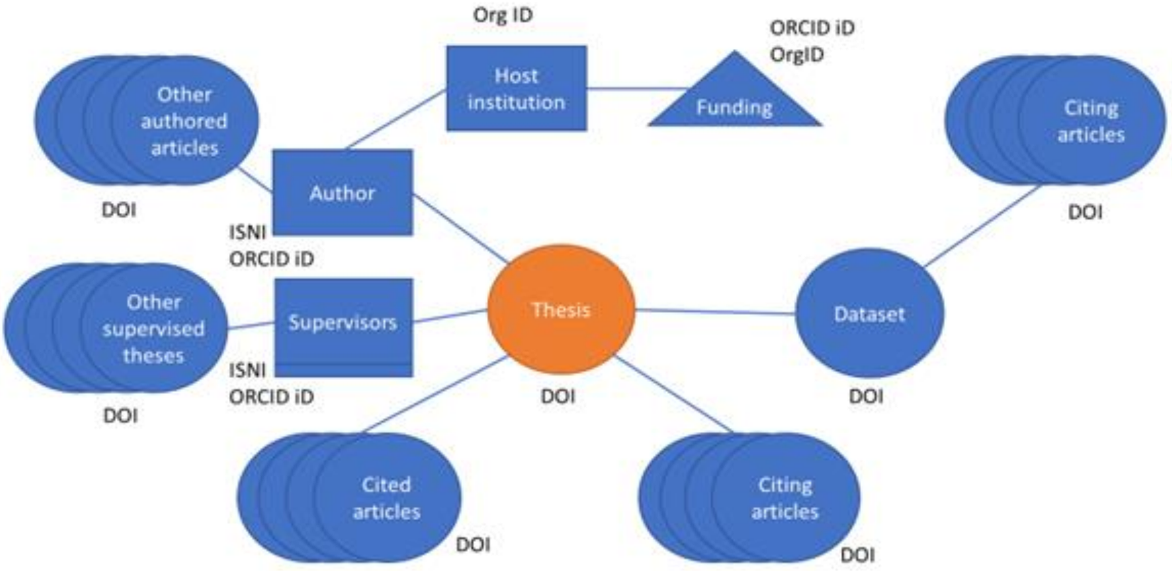


Figure 8: PID Graph representing the potential for metadata within EThOS to be connected by persistent identifiers.

As part of its contribution to the FREYA project, the British Library is working with the Science and Technology Facilities Council (STFC) to link thesis data with identifiers for STFC facilities and equipment. This project is being carried out in two stages. The first with existing “mature” PID types, requiring the registration of STFC’s facilities with an ISNI or Crossref funder DOI. The second with novel “emerging” PID types, presupposing the inclusion of identifiers for funding bodies, facilities, experiments or studies, raw data and associated researchers. For more details about this project, see also Chapter 8 of this report (STFC chapter).

4 European Organization for Nuclear Research (CERN)

4.1 Background and introduction to services

The European Organization for Nuclear Research (CERN)[12] provides particle accelerators, detectors, and the infrastructure for High-Energy Physics (HEP) to produce petabytes of raw data every year[13]. The measurements, or observations, of particle collisions are unrepeatable and inextricably linked to one certain event, making the long time preservation of data a strategic goal (Shiers et al. (2016)). Until reaching internal review, experimental collaborations operate independently and do not share their results and resources, making HEP data not only unique but also sensitive. The workflow of each collaboration is complex and also specific for every experiment. Analyses in HEP are mostly based on three components: data resulting from particle collisions, simulation data and code, which is commonly developed by physicists themselves.

There is a variety of public-facing and restricted services at CERN that have been built to enable and facilitate Open Access, preservation and reproducibility. These services have been developed in the context of Open Science and FAIR data (Wilkinson et al. (2016)) using open source software[14].

In order to ensure the preservation and reusability of HEP physics analyses, the preservation of the whole complex context must be considered. This rather challenging task is carried out by the CERN Analysis Preservation and Reuse framework. This framework mainly consists of two services, CERN Analysis Preservation (CAP)[15] and REusable ANAlyses (REANA)[16]. With these tools, it is possible to describe and capture in a standardized way physics analysis assets - data, software, workflows and computing environment - that are needed to reproduce an analysis even several years after the original scientific results were published. Capturing extensive metadata combined with advanced search capabilities makes CAP an aggregator of high-level physics information about individual physics analyses that facilitates discoverability and reproducibility.

The preservation of physics analyses in CAP ensures long-term accessibility, but does not address how they can be reused, which is a core motivation for researchers to preserve their work in the first place. REANA is a standalone service within this framework that allows researchers to instantiate preserved research data analyses on remote compute clouds using modern container technologies. While REANA can be used to seamlessly rerun analyses preserved in CAP, it can also be used independently to run “live” analyses that are still ongoing and have not yet been published or preserved.

Another data service at CERN that addresses the need for open sharing and promoting reuse is the CERN Open Data portal (COD)[17]. It is an Open Access data repository that contains over 1 petabyte of HEP experimental collision and simulated datasets, software, configuration files, virtual machines, etc. for research and teaching purposes. Data is being published on the platform after embargo periods, which are specified by the data policies of each individual collaboration. Just like in the case of CAP, capturing extensive metadata to describe the complex materials on the portal has been an integral part of the development of this service to ensure preservation, discoverability, and to enable advanced search functionality. The inclusion of usage instructions, related software and other supplementary materials along with the release of data has also proven to be very important and there have already been real-life examples of exploitation (reuse) of the released open content (e.g. Tripathee et al. (2017)).

[12] CERN: https://home.cern/
[13] CERN computing: https://home.cern/about/computing
[14] Invenio Digital Library Framework: https://invenio-software.org/
[15] CERN Analysis Preservation: https://github.com/cernanalysispreservation/analysispreservation.cern.ch
[16] REANA: http://www.reana.io/
[17] CERN Open Data portal: http://opendata.cern.ch/

Finally, CERN also maintains services that are concerned with more "traditional" resources. INSPIRE[18] is the core HEP information system curated by a consortium consisting of DESY, Fermilab, IHEP, SLAC, and CERN. The service focuses on aggregating scholarly works from the global HEP community from various sources ranging from publishers to arXiv. Moreover, INSPIRE also connects to data providers, such as Zenodo, DataVerse and Figshare, and indexes the community data repository HEPData[19]. HEPData contains data points from plots and tables related to several thousands of HEP publications. It is operated by Durham University and CERN. Finally, the CERN Document Server (CDS)[20] is another repository which gives access to CERN works and related HEP scholarly literature (preprints, articles, multimedia, etc.).

4.2 The CERN PID Graph

Currently, PIDs (DataCite DOIs) are minted for data, software, publications and other types of text-based resources (e.g. documentation, guides). Regarding person identifiers, there are ORCID integrations in INSPIRE and HEPData, and ORCID iDs are captured as metadata on CAP and COD. Other persistent identifiers used by CERN services include the INSPIRE ID, which is a unique identifier assigned to all indexed institutions and authors within INSPIRE's institution and author databases respectively, and the analysis ID used internally for physics analyses in CAP.

CERN's PID Graph has been visualized in Figure 9 based on current implementations and established connections.

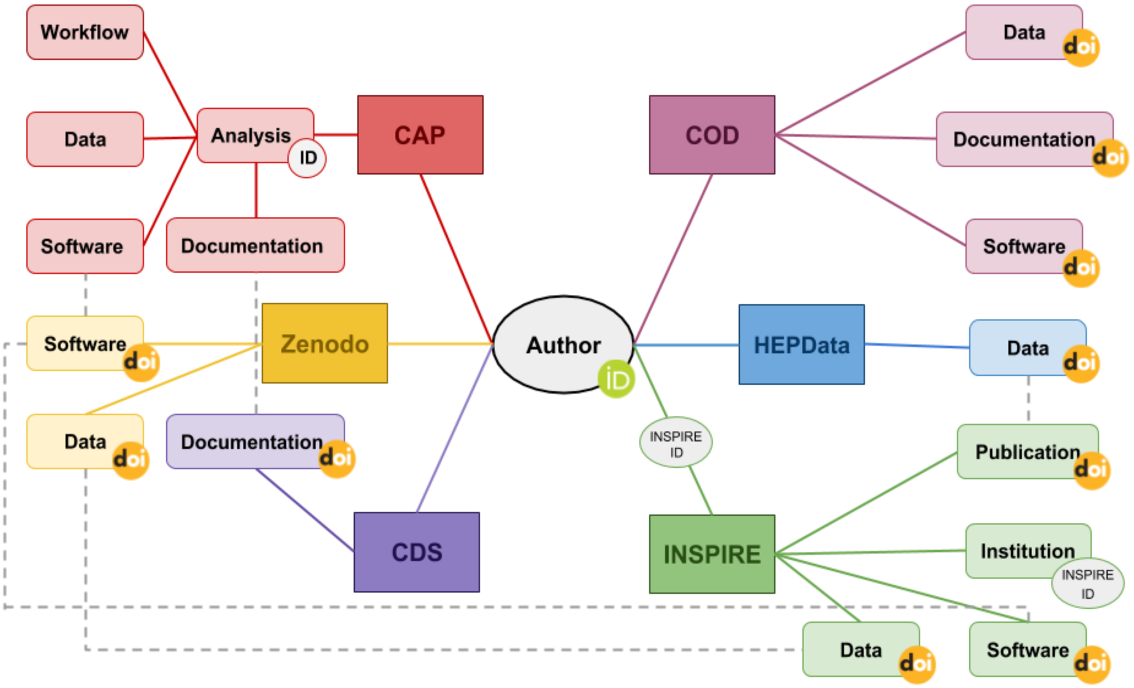


Figure 9: The current CERN PID graph.

In order to enhance and build upon this graph, CERN has been working on implementing a variety of features as part of the FREYA project (and previously THOR[21]), which are described in the following sections.

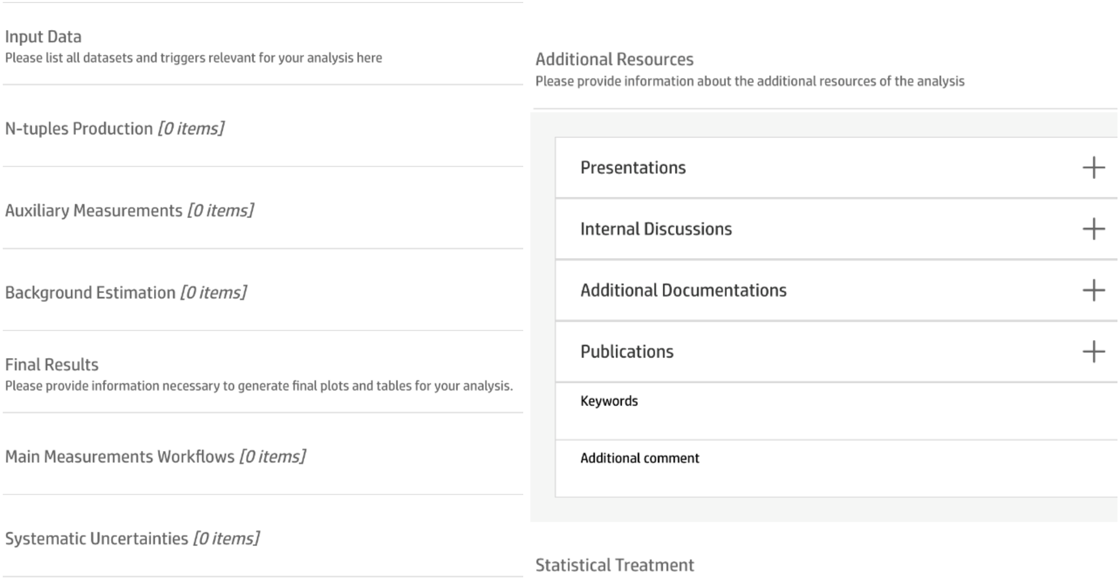
[18] INSPIRE: https://inspirehep.net/
[19] HEPData: https://hepdata.net/
[20] CDS: https://cds.cern.ch/
[21] THOR project: https://project-thor.eu/

4.3 Established integrations: Building the PID Graph

4.3.1 Linking resources

Ensuring that relevant resources are connected with one another is a core principle for all CERN services, as it ensures findability of related resources and that users are aware of the full context of any given resource.

When preserving physics analyses in CERN Analysis Preservation, users are able to include a wide variety of resources relevant to the analysis, i.e. data, software, publications and workflows. CAP is a community PID graph in and of itself as it links all the resources around an analysis (Figure 10).



Input Data
Please list all datasets and triggers relevant for your analysis here

N-tuples Production *[0 items]*

Auxiliary Measurements *[0 items]*

Background Estimation *[0 items]*

Final Results
Please provide information necessary to generate final plots and tables for your analysis.

Main Measurements Workflows *[0 items]*

Systematic Uncertainties *[0 items]*

Additional Resources
Please provide information about the additional resources of the analysis

Presentations +

Internal Discussions +

Additional Documentations +

Publications +

Keywords

Additional comment

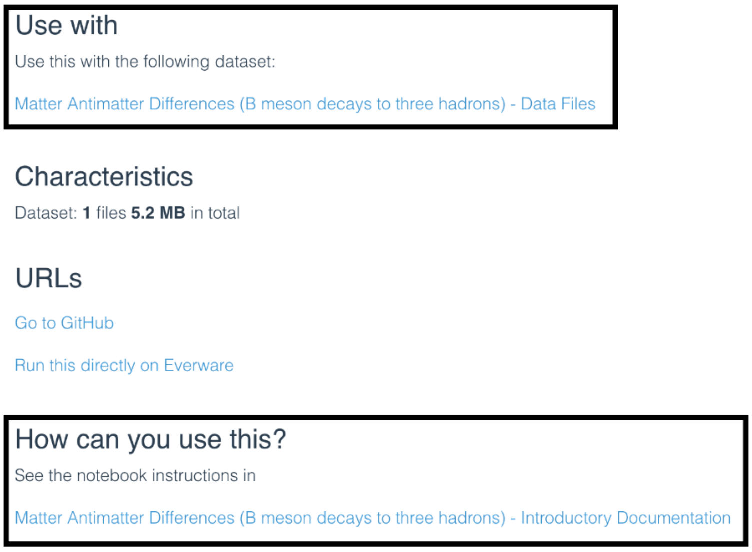
Statistical Treatment

Figure 10: Part of a CAP submission form for physics analyses which gathers various relevant resources.

CAP can be used from the beginning of the research workflow, so the PID Graph can be enriched very early in the research process and not only once research is completed. Connecting resources in a meaningful way was one of the core reasons for the development of the analysis ID concept, which was deployed as part of THOR and FREYA (see more in chapter 4.3.5).

Since CAP has been developed as an internal tool, access restrictions need to be in place for each preserved analysis, so DOIs are not minted. However, PIDs of other resources that are part of an analysis (meaning they are in some way relevant to the analysis) are captured and preserved (e.g. publication PIDs).

In CERN Open Data, there are links connecting software with data and documentation (Figure 11).



Use with

Use this with the following dataset:

Matter Antimatter Differences (B meson decays to three hadrons) - Data Files

Characteristics

Dataset: **1** files **5.2 MB** in total

URLs

Go to GitHub

Run this directly on Everware

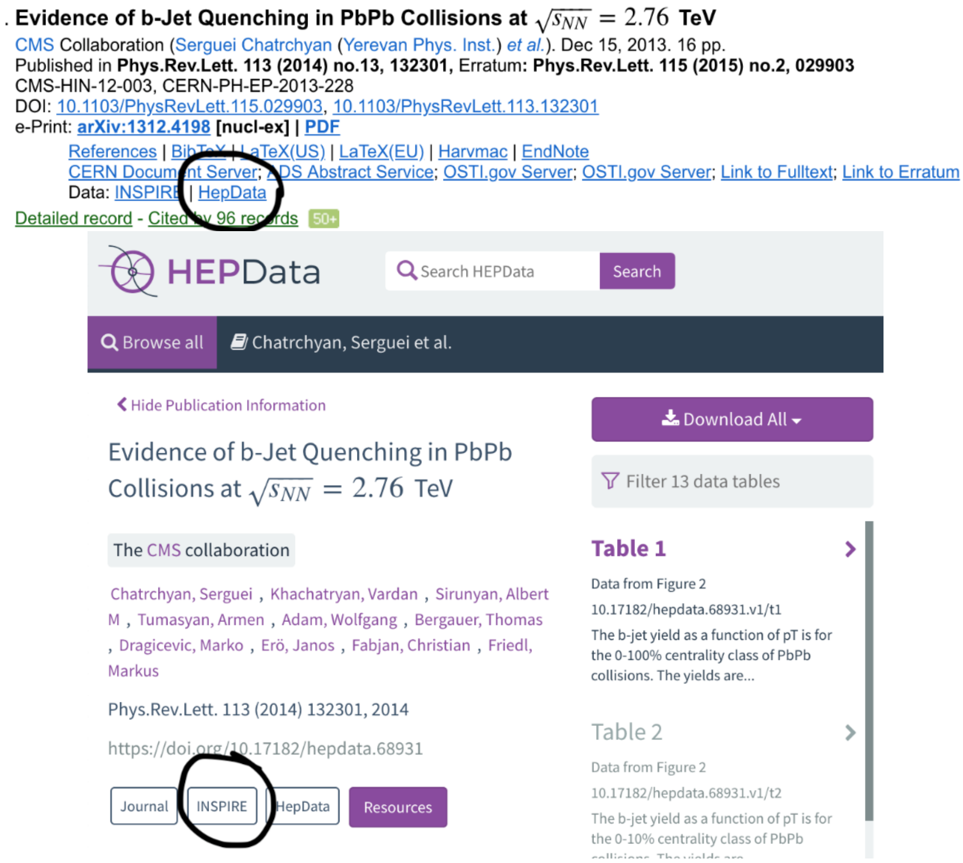
How can you use this?

See the notebook instructions in

Matter Antimatter Differences (B meson decays to three hadrons) - Introductory Documentation

Figure 11: Part of a CERN Open Data software record with links to the corresponding data and documentation records.

INSPIRE and HEPData provide bidirectional links between data and publications (Figure 12).



. **Evidence of b-Jet Quenching in PbPb Collisions at $\sqrt{s_{NN}} = 2.76$ TeV**
CMS Collaboration (Serguei Chatrchyan (Yerevan Phys. Inst.) *et al.*). Dec 15, 2013. 16 pp.
Published in **Phys.Rev.Lett. 113 (2014) no.13, 132301**, Erratum: **Phys.Rev.Lett. 115 (2015) no.2, 029903**
CMS-HIN-12-003, CERN-PH-EP-2013-228
DOI: 10.1103/PhysRevLett.115.029903, 10.1103/PhysRevLett.113.132301
e-Print: **arXiv:1312.4198 [nucl-ex]** | **PDF**

References | BibTeX | LaTeX(US) | LaTeX(EU) | Harvmac | EndNote
CERN Document Server; ADS Abstract Service; OSTI.gov Server; OSTI.gov Server; Link to Fulltext; Link to Erratum
Data: INSPIRE | HepData

Detailed record - Cited by 96 records 50+

Chatrchyan, Serguei et al.

Hide Publication Information

Evidence of b-Jet Quenching in PbPb Collisions at $\sqrt{s_{NN}} = 2.76$ TeV

The CMS collaboration

Chatrchyan, Serguei , Khachatryan, Vardan , Sirunyan, Albert M , Tumasyan, Armen , Adam, Wolfgang , Bergauer, Thomas , Dragicevic, Marko , Erö, Janos , Fabjan, Christian , Friedl, Markus

Phys.Rev.Lett. 113 (2014) 132301, 2014

https://doi.org/10.17182/hepdata.68931

Table 1

Data from Figure 2

10.17182/hepdata.68931.v1/t1

The b-jet yield as a function of pT is for the 0-100% centrality class of PbPb collisions. The yields are...

Table 2

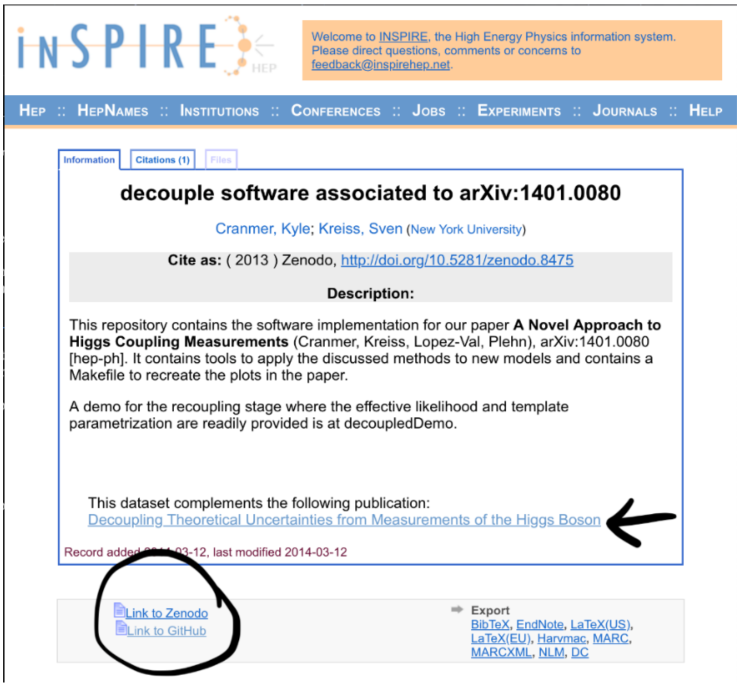
Data from Figure 2

10.17182/hepdata.68931.v1/t2

The b-jet yield as a function of pT is for the 0-10% centrality class of PbPb

Figure 12: Publication record from INSPIRE (top) with a link to the corresponding data on HEPData; the record for the same data on HEPData (bottom) with a link back to the publication on INSPIRE.

There is also GitHub-Zenodo-INSPIRE code integration for software that is related to INSPIRE publications. For such cases, INSPIRE is able to harvest software from GitHub that gets preserved on Zenodo and receives a DOI from there (Figure 13).


iNSPIRE HEP
HEP :: HEPNAMES :: INSTITUTIONS :: CONFERENCES :: JOBS :: EXPERIMENTS :: JOURNALS :: HELP
Information
Citations (1)
Files
Link to Zenodo
Link to GitHub
Export
BibTeX, EndNote, LaTeX(US), LaTeX(EU), Harvmac, MARC, MARCXML, NLM, DC

Welcome to INSPIRE, the High Energy Physics information system. Please direct questions, comments or concerns to feedback@inspirehep.net.

decouple software associated to arXiv:1401.0080

Cranmer, Kyle; Kreiss, Sven (New York University)

Cite as: (2013) Zenodo, http://doi.org/10.5281/zenodo.8475

Description:

This repository contains the software implementation for our paper **A Novel Approach to Higgs Coupling Measurements** (Cranmer, Kreiss, Lopez-Val, Plehn), arXiv:1401.0080 [hep-ph]. It contains tools to apply the discussed methods to new models and contains a Makefile to recreate the plots in the paper.

A demo for the recoupling stage where the effective likelihood and template parametrization are readily provided is at decoupledDemo.

This dataset complements the following publication:
Decoupling Theoretical Uncertainties from Measurements of the Higgs Boson

Record added 2014-03-12, last modified 2014-03-12

Figure 13: INSPIRE software record harvested from Zenodo with links to GitHub and to the corresponding publication.

4.3.2 Data and software citation workflows in HEP

In HEP, there has been a lot of progress in the adoption of data citation practices, which, to a large extent, can be attributed to funders' policies and other mandates. Currently, all data published on CERN platforms get DOIs along with a recommendation about how to cite it and HEPData even makes it possible to cite individual data tables. In the past few years, there has been effort to implement such workflows for software as well.

In the CERN Open Data portal, DOIs are minted for all published software. As already mentioned, CERN Analysis Preservation gathers and preserves a variety of resources that correspond to an individual physics analysis, which also means capturing software PIDs, where they are available. For those cases where a PID is not available for software, an integration with Zenodo has been developed, which makes it possible to generate DOIs for software through the CAP submission form (Figure 14).

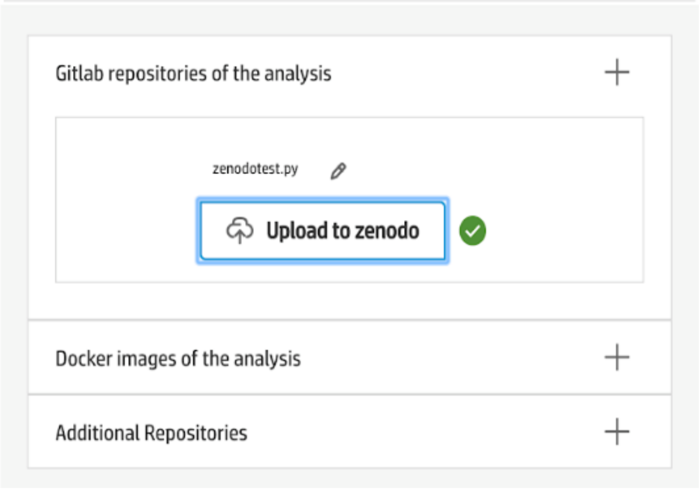
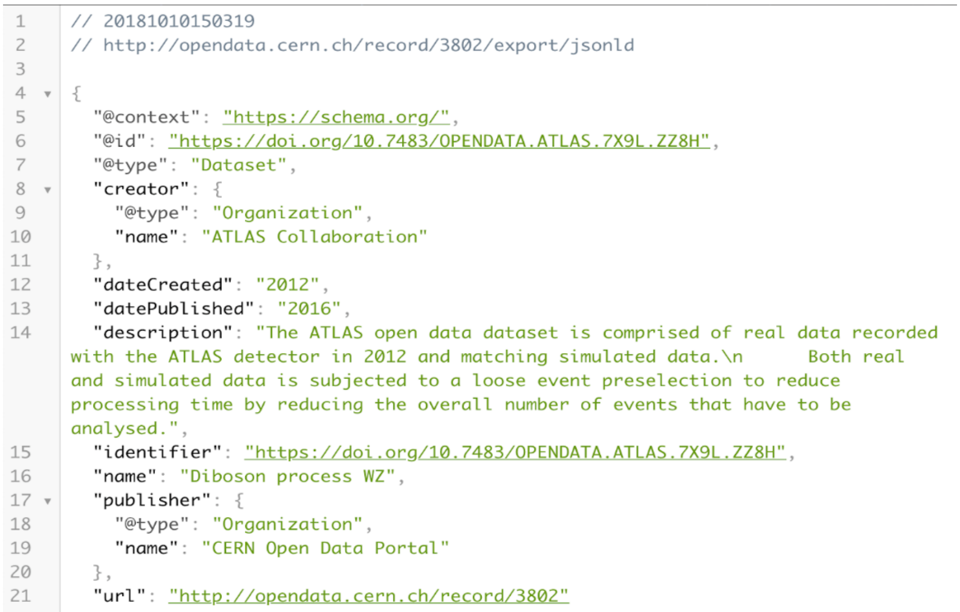

Gitlab repositories of the analysis
+
zenodotest.py
Upload to zenodo
Docker images of the analysis
+
Additional Repositories
+

Figure 14: Zenodo Integration in CAP.

This means that when a user has code that is part of the analysis they are submitting to CAP that needs to be preserved along with the analysis, the code can be fetched from its current location (e.g. GitLab or other platforms). Through the Zenodo integration, the user can then upload said code to Zenodo from the CAP UI and get a DOI registered for it.

4.3.3 Integration of schema.org

The initial version of the schema.org markup using JSON-LD has been released for CERN Open Data and it is currently possible to export records in JSON-LD. In order to expose records in the JSON-LD format, appropriate serializers have been created to transform the data. Record data can be retrieved in the JSON-LD format by specifying the export format in the URL as shown at the very top in Figure 15 below.



```
// 20181010150319
// http://opendata.cern.ch/record/3802/export/jsonld

{
  "@context": "https://schema.org/",
  "@id": "https://doi.org/10.7483/OPENDATA.ATLAS.7X9L.ZZ8H",
  "@type": "Dataset",
  "creator": {
    "@type": "Organization",
    "name": "ATLAS Collaboration"
  },
  "dateCreated": "2012",
  "datePublished": "2016",
  "description": "The ATLAS open data dataset is comprised of real data recorded with the ATLAS detector in 2012 and matching simulated data.\n      Both real and simulated data is subjected to a loose event preselection to reduce processing time by reducing the overall number of events that have to be analysed.",
  "identifier": "https://doi.org/10.7483/OPENDATA.ATLAS.7X9L.ZZ8H",
  "name": "Diboson process WZ",
  "publisher": {
    "@type": "Organization",
    "name": "CERN Open Data Portal"
  },
  "url": "http://opendata.cern.ch/record/3802"
```

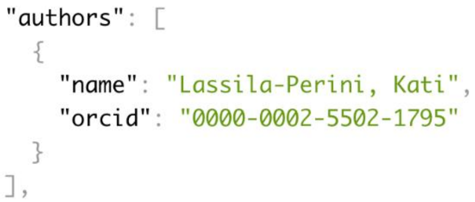
Figure 15: Schema.org-JSON-LD markup in a CERN Open Data dataset record.

The schema.org implementation in CERN Open Data is expected to be extended in the future once solutions are found for community-specific challenges, such as the best way to expose dataset records that contain several thousands of files (“DataDownload” fields).

While HEPData is also marked up with schema.org using Microdata, it is envisioned that the encoding will be changed to JSON-LD.

4.3.4 ORCID Integration

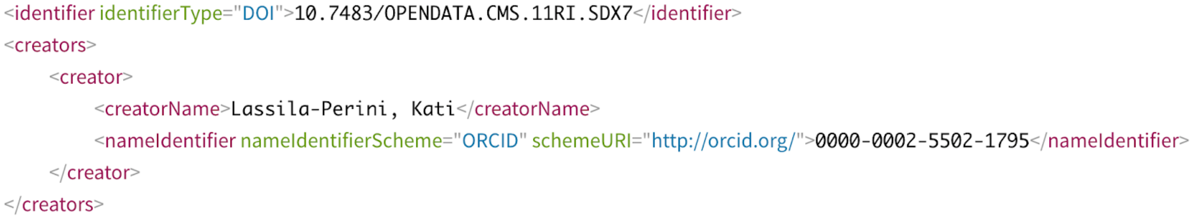
Each of our services provides a way to uniquely identify authors and contributors by integrating ORCID iDs. All resources released in the CERN Open Data portal refer to an author (collaboration or person); for people, person PIDs can also be captured in the metadata, such as an ORCID iD (Figure 16).



```
"authors": [
  {
    "name": "Lassila-Perini, Kati",
    "orcid": "0000-0002-5502-1795"
  }
],
```

Figure 16: Metadata of a CERN Open Data record that includes an ORCID iD.

For records on the portal that have ORCID iDs as well as DOIs, the ORCID iDs also get included in the DOI metadata, as shown in Figure 17.



```
<identifier identifierType="DOI">10.7483/OPENDATA.CMS.11RI.SDX7</identifier>
<creators>
    <creator>
        <creatorName>Lassila-Perini, Kati</creatorName>
        <nameIdentifier nameIdentifierScheme="ORCID" schemeURI="http://orcid.org/">0000-0002-5502-1795</nameIdentifier>
    </creator>
</creators>
```

Figure 17: Metadata of a CERN Open Data DOI that includes and ORCID iD.

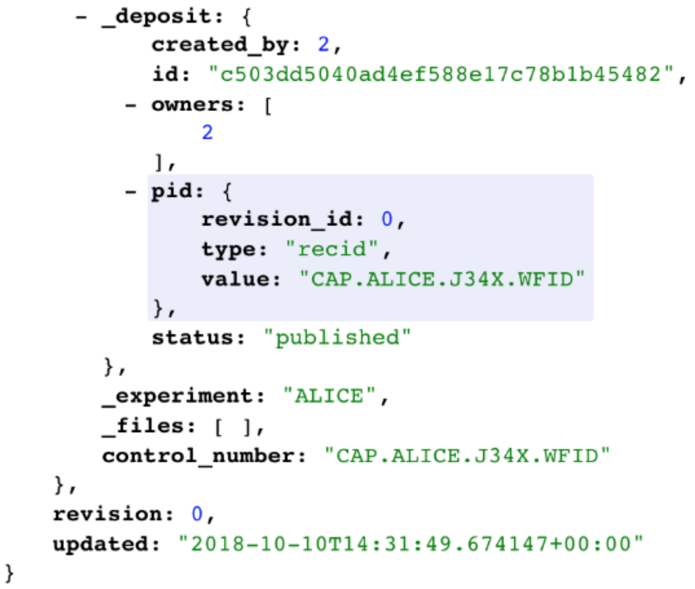
CERN Analysis Preservation does this automatically for users. When inputting a name of an author, collaborator or reviewer the user will get a suggestion to include their ORCID iD, if one was already registered for the name being entered.

There is also ORCID integration as part of INSPIRE's author profiles (HEPNames), which is a directory for HEP authors that links them to their research outputs (within INSPIRE), while also providing relevant metrics. This includes logging in using an ORCID iD, which is also possible in HEPData, and connecting an ORCID iD to an INSPIRE author profile.

4.3.5 The analysis ID

Handling such complex and rich analyses consisting of dozens of permanently evolving components from different sources underlines the need for PIDs in order to refer to an entire analysis or versions of an analysis.

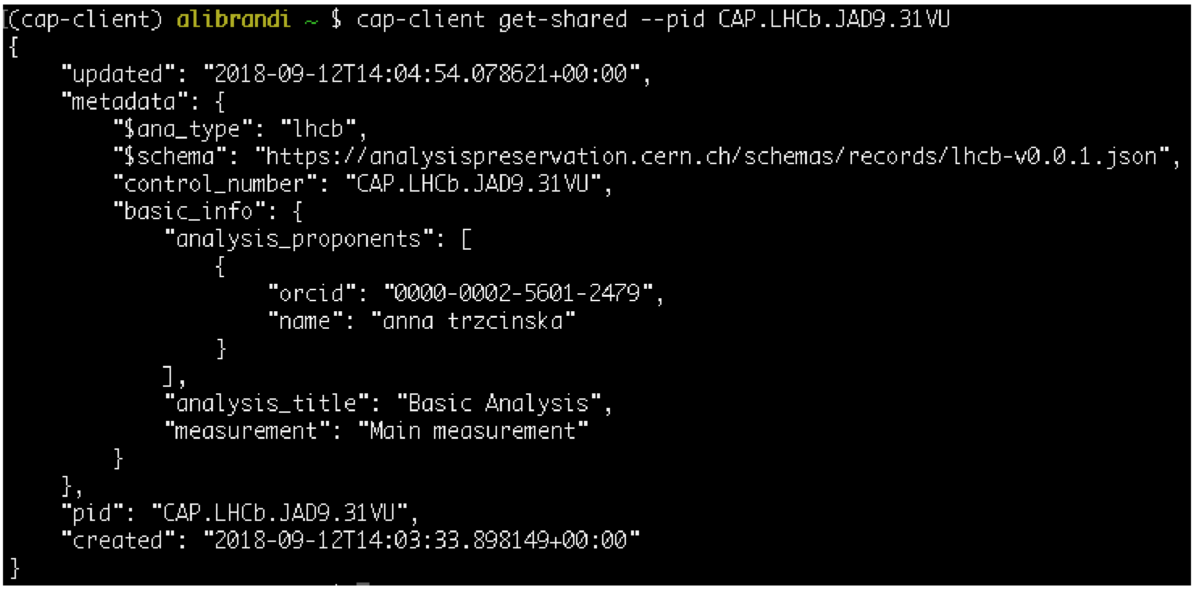
As already mentioned, CAP is a restricted-access platform and as a result it cannot include external DOIs because there are no public landing pages available for the DOIs to resolve to. Therefore, we developed the analysis ID, which is a UUID-based unique identifier for analyses preserved in CAP. As shown in Figure 18, slugs are also in place for these UUIDs in order for them to have a more meaningful display format.



```
          - _deposit: {
                  created_by: 2,
                  id: "c503dd5040ad4ef588e17c78b1b45482",
                - owners: [
                      2
                  ],
                - pid: {
                      revision_id: 0,
                      type: "recid",
                      value: "CAP.ALICE.J34X.WFID"
                  },
                  status: "published"
              },
              _experiment: "ALICE",
              _files: [ ],
              control_number: "CAP.ALICE.J34X.WFID"
          },
          revision: 0,
          updated: "2018-10-10T14:31:49.674147+00:00"
      }
```

Figure 18: The analysis ID in the CAP service.

Within CAP, this analysis ID is used to carry out a variety of functions via the command line client (Figure 19). For example, it is possible to edit metadata or upload a file to an analysis with a given PID or to retrieve, delete, publish it, etc.



```
(cap-client) alibrandi ~ $ cap-client get-shared --pid CAP.LHCb.JAD9.31VU
{
    "updated": "2018-09-12T14:04:54.078621+00:00",
    "metadata": {
        "$ana_type": "lhcb",
        "$schema": "https://analysispreservation.cern.ch/schemas/records/lhcb-v0.0.1.json",
        "control_number": "CAP.LHCb.JAD9.31VU",
        "basic_info": {
            "analysis_proponents": [
                {
                    "orcid": "0000-0002-5601-2479",
                    "name": "anna trzcinska"
                }
            ],
            "analysis_title": "Basic Analysis",
            "measurement": "Main measurement"
        }
    },
    "pid": "CAP.LHCb.JAD9.31VU",
    "created": "2018-09-12T14:03:33.898149+00:00"
}
```

Figure 19: Use of the analysis ID as part of executing commands related to analyses through CAP's command-line tool.

4.4 Integrating new PID types

When considering CERN's services overall, it is evident that some of the emerging PID types can certainly contribute to a more enhanced PID Graph that adds more value. The consideration of these new PID types differentiates two scenarios: integration of new PIDs to replace existing internal IDs and adopting new PID types to identify resources that had not been previously implemented at all in any of the services. It is important to note that the potential integration of either of the two is guided by the evolution of the services, the maturity and suitability of the emerging PID type, and whether there is a need for said new PID types.

The following emerging PIDs are considered relevant for CERN services at this time:

- Organizations: The INSPIRE service includes an Institutions and Experiments database. The first could be an appropriate use case for the emerging PID for organizations (see Deliverable 3.1). A collaboration with the Research Organization Registry Community (ROR)[22] could be a possibility and needs to be investigated. One possible issue for this integration could be the granularity of the entities in the registry, as INSPIRE is a manually curated service and includes many organization identifiers that have been getting added for decades.
- Instruments: In the HEP community there are instruments - used to measure phenomena or gather data - and experiments, a term that usually refers to the project or collaboration running or operating an instrument. At CERN, there is a wide range of ongoing experiments/instruments and with that also a wide range of related databases. The Experiments database on INSPIRE lists experiments within the HEP community. Another relevant database is Grey Book[23], which includes basic metadata about experiments at CERN. The forthcoming first schemas of PIDInst[24] (RDA Working Group) will need to be tested thoroughly for possible suitability.
- Grants: For some of the platforms, i.e. INSPIRE and potentially CERN Analysis Preservation, it might be interesting to include identifiers for grants. Looking at it from the PID Graph perspective, it is expected to be of interest to users to be able to have established connections between authors, grants, publications, data and software. It could be used, for example, for grant reporting or job

[22] Research Organization Registry: https://www.ror.community/
[23] Grey Book database: https://greybook.cern.ch/greybook/experiment/list
[24] Persistent Identification of Instruments WG: https://www.rd-alliance.org/groups/persistent-identification-instruments-wg

applications. However, it requires wide enough coverage of grants that are used by the HEP community for the users to be able to benefit from it. This will need further investigation as the grant ID standards develop further.

- Workflows: As indicated in Deliverable 3.1, some platforms in the Life Sciences already use PIDs (e.g. UUIDs) for their workflows. Looking at the use case of the aforementioned REANA project, one could imaging having a DOI for different "types" of analyses, if they were to be published. A workflow PID could be given, for example, to a preserved analysis that is reproduced using the same workflow configuration or to an analysis that is being rerun with changed parameters, which means that the inputs are different and consequently so are the outputs. Overall, such citation workflows would not be any different to data or software citation. However, the implementation would need to be agreed upon with the community.
- Conferences: CERN offers Indico[25], an Open Source framework for conferences, which is widely used within the community and beyond. The emerging discussions about having a PID for conferences should include such software packages. A discussion about this has been started and needs to be continued, but it is not expected that this will be delivered in the timeline of this project.

Based on the above, potential connections that could be established between these new PIDs and the existing CERN services have been visualized in Figure 20 below.

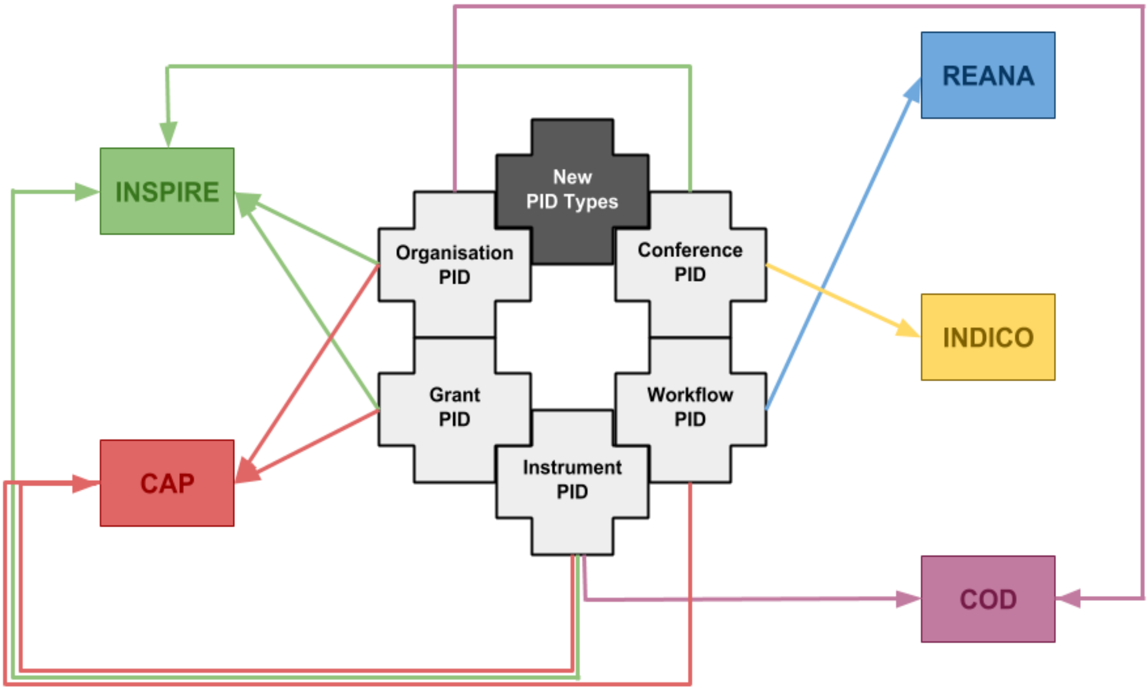


Figure 20: Possible new puzzle pieces that could be connected to the existing CERN PID Graph in the future.

[25] Indico: https://indico.cern.ch/

5 Data Archiving and Networked Services (KNAW–DANS)

5.1 Introduction to services

DANS (Data Archiving and Networked Services)[26] is the Netherlands Institute for permanent access to digital research resources. It promotes sustained access to digital research data and encourages researchers to make their digital research data and related outputs Findable, Accessible, Interoperable and Reusable (in accordance with the FAIR principles). DANS' pilot application involves two of its core services: the research information service NARCIS and the EASY repository.

NARCIS (National Academic Research and Collaborations Information System)[27] is the national portal for information about researchers and their work. The portal provides access to scientific information, including open and restricted-access publications from the repositories of all Dutch universities and various other national research organizations. It also includes scientific datasets from several archives, including DANS' own long-term archive, EASY[28], and descriptions of research projects, researchers and research institutes. NARCIS is used by a wide range of users, including researchers, policy makers, educators, journalists and the general public.

EASY is an online archiving system with tens of thousands of datasets from research completed by researchers and institutes. Although EASY can store data from all disciplines, the majority of the datasets it holds are from the Humanities and Social Sciences. EASY is also used as a long-term preservation system for services like Mendeley and Dryad. In total, it contains approximately 75.000 datasets primarily from the Humanities and Social Sciences.

5.2 The NARCIS PID Graph

NARCIS indexes more than 2 million objects, often with more than one PID. About 40 scientific organizations provide most of the information in NARCIS through OAI-PMH. Within the Dutch national infrastructure, all scientific organizations use common metadata formats - either MODS or DataCite metadata - for the exchange of information with NARCIS and other national services. NARCIS harvests these services, aggregates and standardizes the metadata, and indexes all available PIDs: DOIs, URNs Handles, PubMed IDs, Web of Science IDs, etc. For persons, NARCIS supports ORCID iDs, ISNIs and DAIs (Digital Author Identifier).

NARCIS harvests five different content types: publications (1.84 million), datasets (270.000), research and funding projects (70.000), researchers (60.000) and research organizations (3.000). NARCIS creates bi-directional links between publications and data and shows the object type (article, thesis, datasets, etc.), and the title of the item linked. NARCIS also acts as a repository itself by offering an API for other content services, like Google Scholar, OpenAIRE, Dart Europe E-Thesis portal, EBSCO, ProQuest and others. Figure 21 below gives a schematic overview of the NARCIS PID Graph.

[26] DANS: https://dans.knaw.nl/en
[27] NARCIS: http://www.narcis.nl
[28] EASY: http://www.easy.dans.knaw.nl

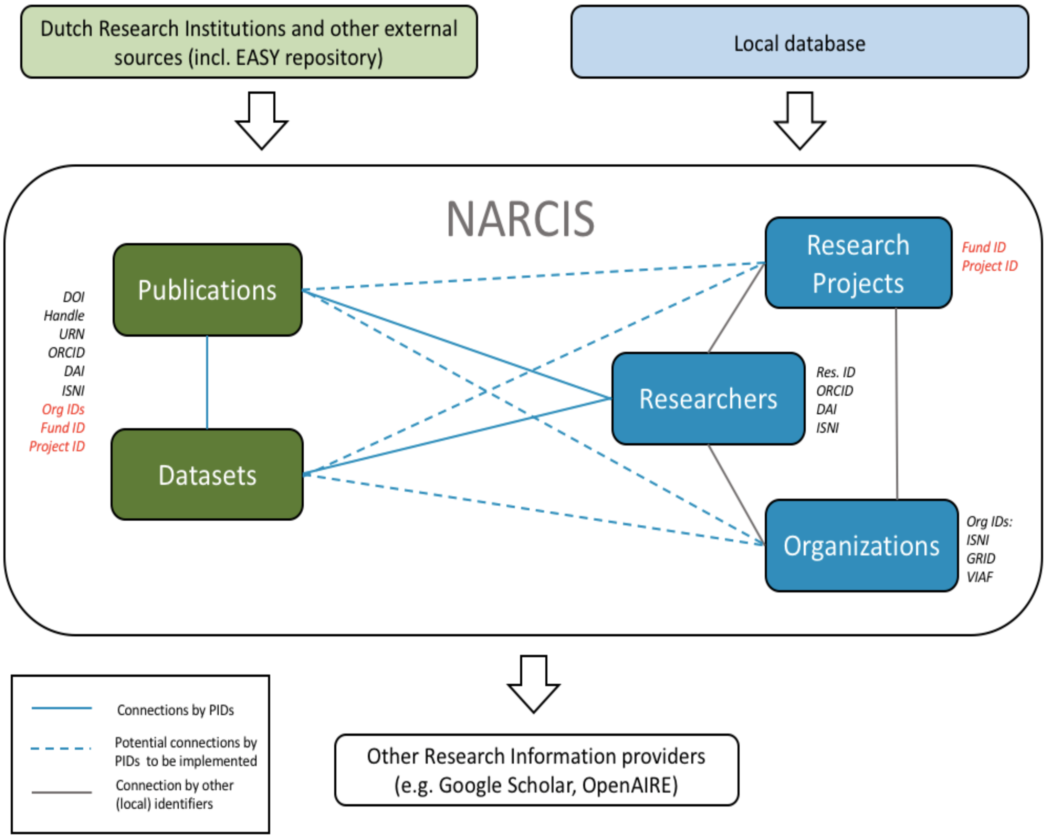

Dutch Research Institutions and other external sources (incl. EASY repository)
Local database
NARCIS
Publications
Datasets
Research Projects
Researchers
Organizations
DOI
Handle
URN
ORCID
DAI
ISNI
Org IDs
Fund ID
Project ID
Fund ID
Project ID
Res. ID
ORCID
DAI
ISNI
Org IDs:
ISNI
GRID
VIAF
Connections by PIDs
Potential connections by PIDs to be implemented
Connection by other (local) identifiers
Other Research Information providers (e.g. Google Scholar, OpenAIRE)

Figure 21: The NARCIS PID Graph: a schematic overview of the main information types, their sources, and their connections via PIDs in the metadata. Blue solid lines indicate connections via PIDs, grey solid lines indicate connections via other (local) identifiers. Blue dashed lines show potential connections using PIDs to be implemented in the future (marked in red).

In order to integrate and interconnect mature PIDs within the NARCIS PID Graph, DANS has worked on developing several features as part of its tasks for this Work Package, which are described in the following sections.

5.3 Building the PID Graph: Mature PID types

5.3.1 ORCID integration

During the first year of the FREYA project, more than 7.500 ORCID iDs were added to NARCIS. By the end of 2018, NARCIS will show for each researcher which publications are harvested directly by NARCIS from Dutch institutions and which ones are linked to their ORCID, which NARCIS then extracts from ORCID's public API.

This will result in two different scenarios from a viewpoint of an individual researcher:

- New connections between a researcher and their scientific output (publications or data) that is already in NARCIS. NARCIS takes information about the relations between researchers and objects form ORCID and adds this information to the portal.
- Scientific output not available in NARCIS that can be added to the portal. NARCIS compares the PIDs of new objects with the PIDs of objects stored in ORCID and is able to differentiate between truly new objects and objects already part of NARCIS.

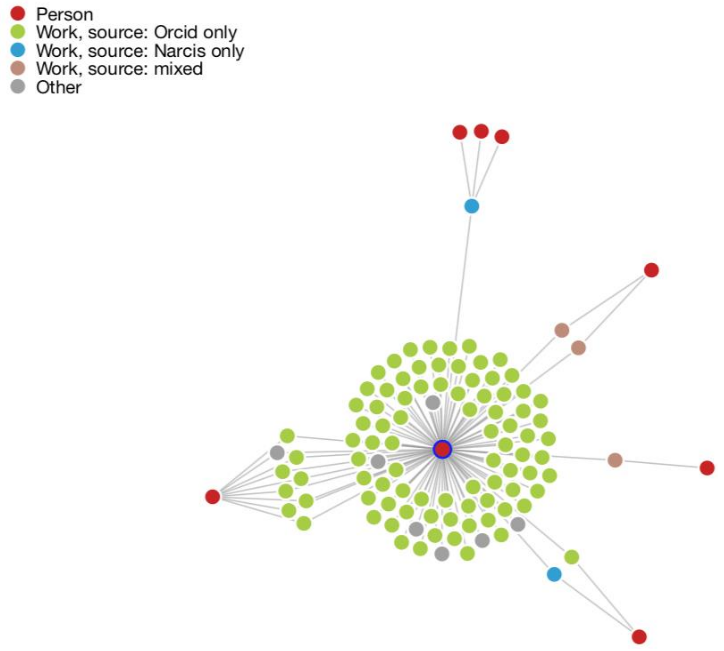


Figure 22: Visualization of a PID Graph for objects in NARCIS vs in ORCID produced by an individual researcher.

In the scenario shown in Figure 22, the researcher has around 100 publications in ORCID and five in NARCIS, three of which are present in both portals. This shows big differences between the information harvested from the Dutch institutions and those in ORCID, illustrating that adding ORCID iDs can have a considerable impact on the completion of researcher records. Generally, these proportions can show a lot of variation, ranging from all publications being stored only in NARCIS or only in ORCID.

In addition to integrating more information and relations, we are working on a visual feature that will show the origin of the objects for persons with an ORCID iD in the NARCIS portal. As the number of researchers with an ORCID iD grows, NARCIS will be able show a PID Graph for all its researchers. Other information services such as DataCite, Crossref or Web of Science can also be connected.

5.3.2 Integration of schema.org

Further integrating PIDs in the EASY and NARCIS services and publishing metadata via schema.org makes the content more compatible and interoperable with national and international scientific information systems and Open Science platforms.

Within the FREYA project the majority of partners have already managed to expose at least some basic metadata for a given resource using schema.org. In the first year of FREYA, schema.org was implemented in NARCIS for publications, datasets, researchers and organizations.

In the NARCIS release of September 2018, schema.org became available. More than 1.8 million publications, 271.000 datasets, 58.000 researchers and 3.000 research organizations are marked up in schema.org, providing PIDs (where available) of objects, researchers and organizations to other services.

5.3.3 Unpaywall integration

DANS promotes Open Access and supports NARCIS users in finding Open Access publications. As part of this, we have established an integration with Unpaywall[29], a service that harvests full-text articles from repositories all over the world. However, when comparing the access rights for publications in NARCIS (left

[29] Unpaywall: https://unpaywall.org/

in Figure 23) to the access rights for the same publications shown by Unpaywall (right), it is apparent that there are a lot of discrepancies between the two sources.

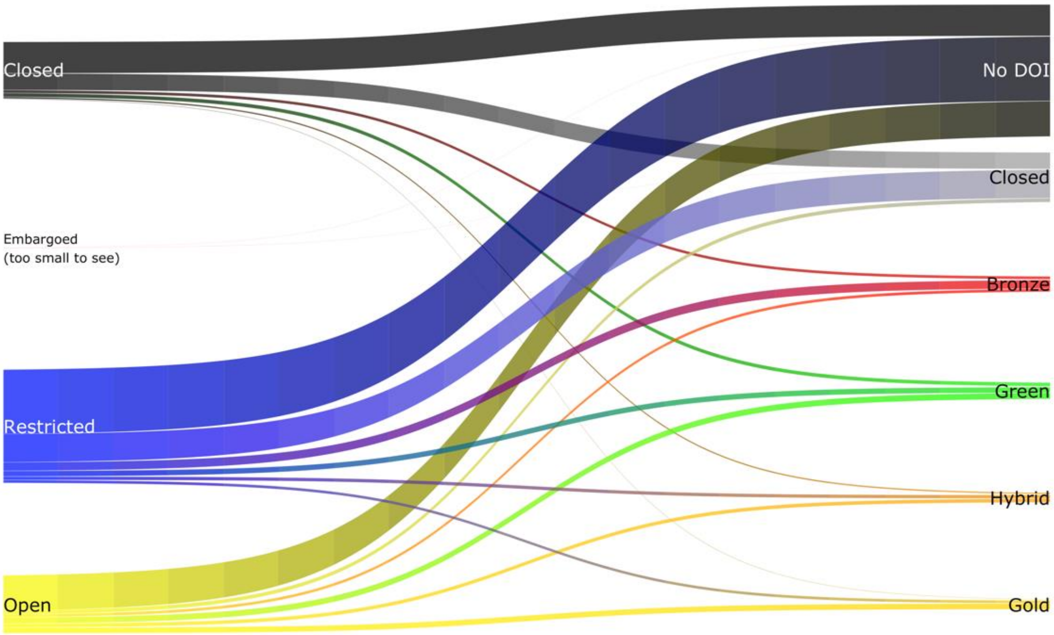


Figure 23: Publication access rights according to NARCIS (left) compared to Unpaywall (right).

The figure shows that in many cases freely-accessible NARCIS publications are listed as being closed or restricted by Unpaywall.

DANS has also made an integration to support NARCIS users in finding Open Access publications by giving them access to alternative locations offered by Unpaywall. NARCIS presents the Unpaywall logo next to every DOI in the system with a short explanation. By clicking on the logo or the accompanying link, NARCIS will retrieve all Unpaywall information belonging to a DOI and present this in the NARCIS portal. Unpaywall offers an API where information can be retrieved using DOIs. This integration is of great benefit to NARCIS users and at the same time it supports Open Access.

5.3.4 Article-data linking

DANS has implemented the necessary steps to create links between publications and data in collaboration with three NARCIS partners and the EASY archive. We used real use cases from NARCIS content providers to help us understand what universities and other research organizations want to see realized. Although these four organizations do not represent the entire scientific community, they are concrete use cases through which we can better understand the requirements for such integrations.

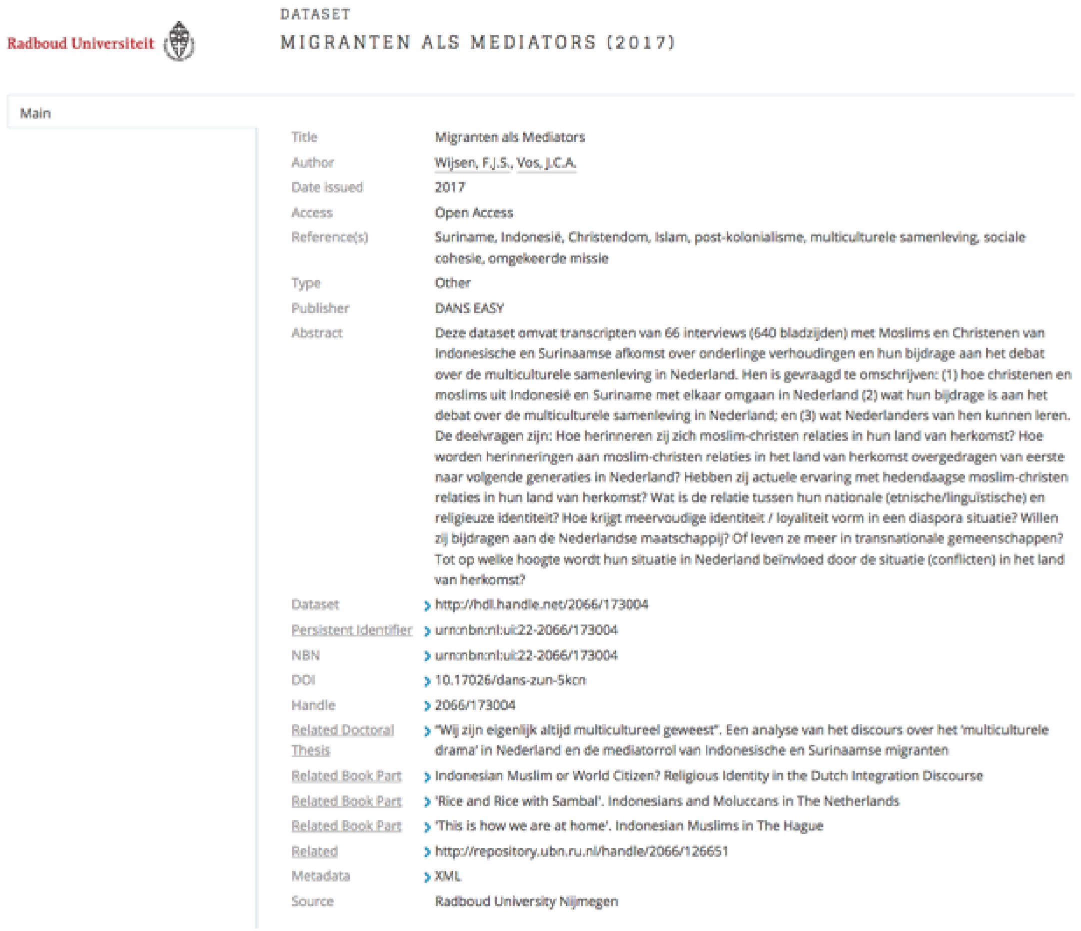
In this case, it was important to implement something that works on a general level. NARCIS contains information from more than 45 different repositories and services and these integrations can be seen as the starting point for implementing article-data linking practices in other repositories as well. Other than linking publications to datasets, NARCIS also provides links between publications or between datasets. One of the benefits of article-data linking, or linking in general, is that it provides more context to research.

5.3.5 Use cases

5.3.5.1 Use Case: Radboud University

Radboud University, one of the largest Dutch universities, recently started to collect datasets and has published around 200 datasets in NARCIS so far. For Radboud University, one of the key conditions was to establish links to related publications and other datasets to give more context to these publications and datasets.

Figure 24 below shows a record of a dataset in NARCIS.



Radboud Universiteit

DATASET

MIGRANTEN ALS MEDIATORS (2017)

Main

Title	Migranten als Mediators
Author	Wijsen, F.J.S., Vos, J.C.A.
Date issued	2017
Access	Open Access
Reference(s)	Suriname, Indonesië, Christendom, Islam, post-kolonialisme, multiculturele samenleving, sociale cohesie, omgekeerde missie
Type	Other
Publisher	DANS EASY
Abstract	Deze dataset omvat transcripten van 66 interviews (640 bladzijden) met Moslims en Christenen van Indonesische en Surinaamse afkomst over onderlinge verhoudingen en hun bijdrage aan het debat over de multiculturele samenleving in Nederland. Hen is gevraagd te omschrijven: (1) hoe christenen en moslims uit Indonesië en Suriname met elkaar omgaan in Nederland (2) wat hun bijdrage is aan het debat over de multiculturele samenleving in Nederland; en (3) wat Nederlanders van hen kunnen leren. De deelvragen zijn: Hoe herinneren zij zich moslim-christen relaties in hun land van herkomst? Hoe worden herinneringen aan moslim-christen relaties in het land van herkomst overgedragen van eerste naar volgende generaties in Nederland? Hebben zij actuele ervaring met hedendaagse moslim-christen relaties in hun land van herkomst? Wat is de relatie tussen hun nationale (etnische/linguïstische) en religieuze identiteit? Hoe krijgt meervoudige identiteit / loyaliteit vorm in een diaspora situatie? Willen zij bijdragen aan de Nederlandse maatschappij? Of leven ze meer in transnationale gemeenschappen? Tot op welke hoogte wordt hun situatie in Nederland beïnvloed door de situatie (conflicten) in het land van herkomst?
Dataset	› http://hdl.handle.net/2066/173004
Persistent Identifier	› urn:nbn:nl:ui:22-2066/173004
NBN	› urn:nbn:nl:ui:22-2066/173004
DOI	› 10.17026/dans-zun-5kcn
Handle	› 2066/173004
Related Doctoral Thesis	› "Wij zijn eigenlijk altijd multicultureel geweest". Een analyse van het discours over het 'multiculturele drama' in Nederland en de mediatorrol van Indonesische en Surinaamse migranten
Related Book Part	› Indonesian Muslim or World Citizen? Religious Identity in the Dutch Integration Discourse
Related Book Part	› 'Rice and Rice with Sambal'. Indonesians and Moluccans in The Netherlands
Related Book Part	› 'This is how we are at home'. Indonesian Muslims in The Hague
Related	› http://repository.ubn.ru.nl/handle/2066/126651
Metadata	› XML
Source	Radboud University Nijmegen

Figure 24: Screenshot of a full record of a dataset in NARCIS.

Figure 25 shows the PID Graph for the same dataset. The data was used for a doctoral thesis and the graph shows the connections between the data and the thesis, the contributors and three related book chapters, all connected through PIDs.

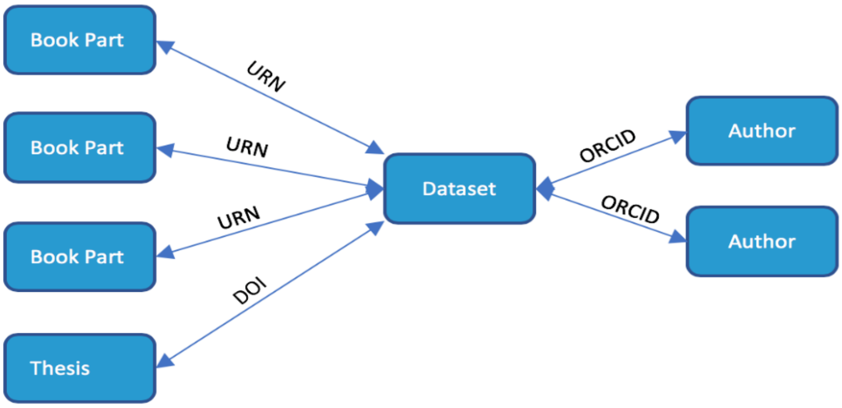


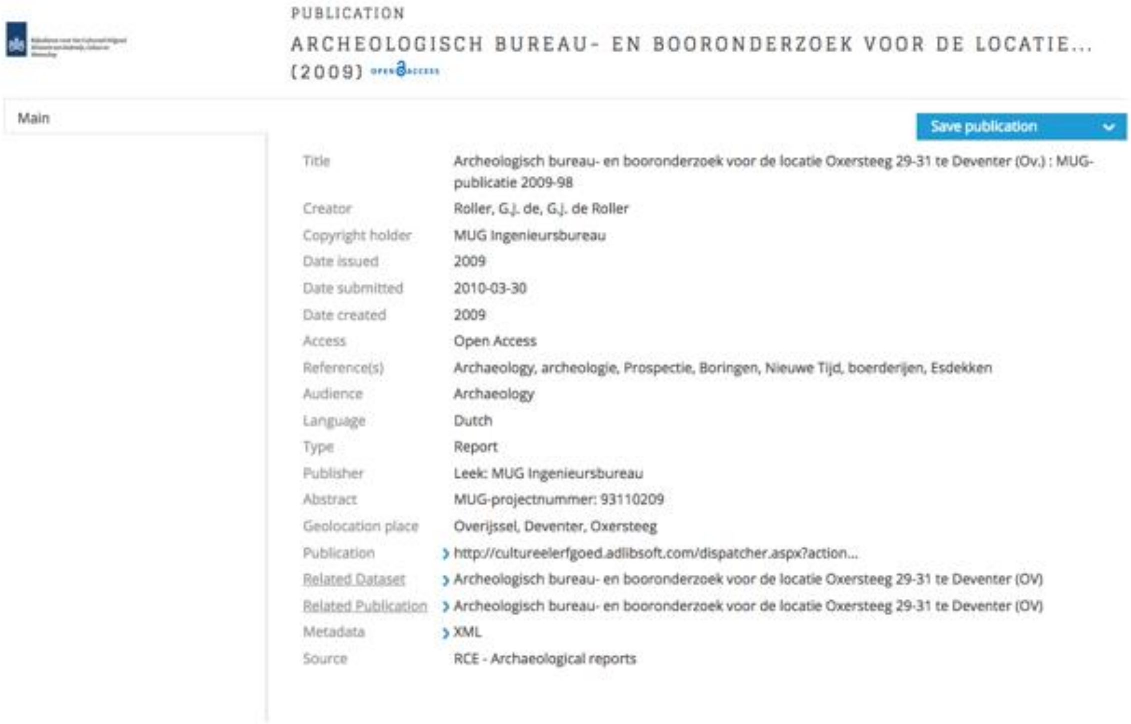
Figure 25: Graph of the Radboud University dataset in NARCIS. All objects have PIDs and give direct access to related resources.

5.3.5.2 Use Case: Dutch Cultural Heritage Agency (RCE)

The Dutch Cultural Heritage Agency (RCE) owns a collection of more than 100.000 archaeological reports. EASY, which holds the E-depot for Dutch Archaeology (EDNA), contains more than 30.000 archaeological datasets. Since 2007, archaeologists in the Netherlands are obligated to deposit their data in EASY. The RCE makes the metadata of their collection available through NARCIS to the archaeological community.

It would be a great asset to archaeological researchers if NARCIS could link these reports to the data in EASY and vice versa. So far, we have been able to link more than 7.000 reports to their corresponding datasets in EASY and plans are made to add links to more or, if possible, to all archeological datasets.

Figure 26 shows a record for such an archaeological report in NARCIS. A link to the dataset and any other relations have been added to the record, linked using PIDs.



ARCHEOLOGISCH BUREAU- EN BOORONDERZOEK VOOR DE LOCATIE... (2009)

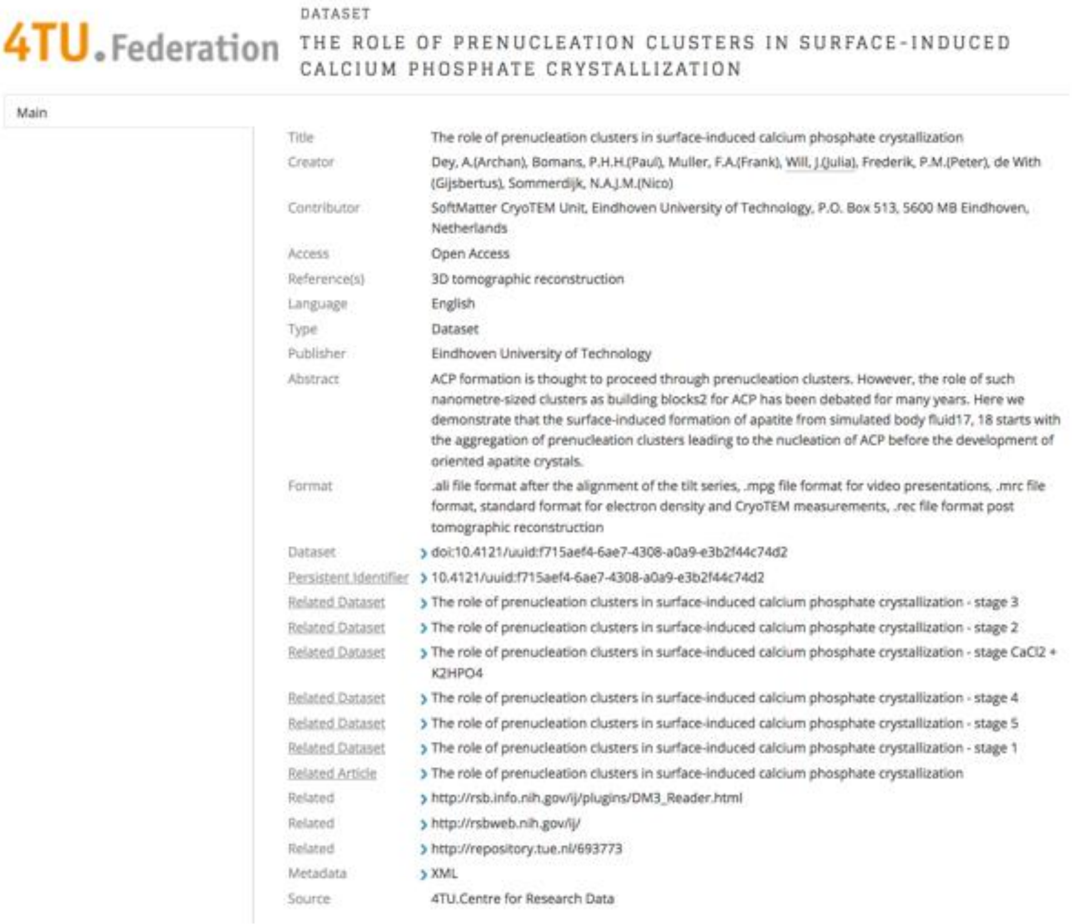
Title	Archeologisch bureau- en booronderzoek voor de locatie Oxersteeg 29-31 te Deventer (Ov.) : MUG-publicatie 2009-98
Creator	Roller, G.J. de, G.J. de Roller
Copyright holder	MUG Ingenieursbureau
Date issued	2009
Date submitted	2010-03-30
Date created	2009
Access	Open Access
Reference(s)	Archaeology, archeologie, Prospectie, Boringen, Nieuwe Tijd, boerderijen, Esdekken
Audience	Archaeology
Language	Dutch
Type	Report
Publisher	Leek: MUG Ingenieursbureau
Abstract	MUG-projectnummer: 93110209
Geolocation place	Overijssel, Deventer, Oxersteeg
Publication	› http://cultureelerfgoed.adlibsoft.com/dispatcher.aspx?action...
Related Dataset	› Archeologisch bureau- en booronderzoek voor de locatie Oxersteeg 29-31 te Deventer (OV)
Related Publication	› Archeologisch bureau- en booronderzoek voor de locatie Oxersteeg 29-31 te Deventer (OV)
Metadata	› XML
Source	RCE - Archaeological reports

Figure 26: Screenshot of a full record of an archaeological report in NARCIS.

5.3.5.3 Use Case: 4TU Federation data archive

The 4TU Federation consists of the four universities of technology in the Netherlands: Technical University Delft, Eindhoven University of Technology, University of Twente and University of Wageningen. The 4TU Federation data archive consists of almost 4.500 datasets, of which the metadata is aggregated into NARCIS. Many datasets contain relations to other resources: datasets, other versions of datasets, publications or websites.

NARCIS supports the interlinkage by adding links to referred objects. This is also done for objects outside the 4TU data archive. Figure 27 shows the full record of a 4TU dataset in NARCIS.



4TU.Federation

DATASET

THE ROLE OF PRENUCLEATION CLUSTERS IN SURFACE-INDUCED CALCIUM PHOSPHATE CRYSTALLIZATION

Main

Title	The role of prenucleation clusters in surface-induced calcium phosphate crystallization
Creator	Dey, A.(Archan), Bomans, P.H.H.(Paul), Muller, F.A.(Frank), Will, J.(Julia), Frederik, P.M.(Peter), de With (Gijsbertus), Sommerdijk, N.A.J.M.(Nico)
Contributor	SoftMatter CryoTEM Unit, Eindhoven University of Technology, P.O. Box 513, 5600 MB Eindhoven, Netherlands
Access	Open Access
Reference(s)	3D tomographic reconstruction
Language	English
Type	Dataset
Publisher	Eindhoven University of Technology
Abstract	ACP formation is thought to proceed through prenucleation clusters. However, the role of such nanometre-sized clusters as building blocks2 for ACP has been debated for many years. Here we demonstrate that the surface-induced formation of apatite from simulated body fluid17, 18 starts with the aggregation of prenucleation clusters leading to the nucleation of ACP before the development of oriented apatite crystals.
Format	.ali file format after the alignment of the tilt series, .mpg file format for video presentations, .mrc file format, standard format for electron density and CryoTEM measurements, .rec file format post tomographic reconstruction
Dataset	› doi:10.4121/uuid:f715aef4-6ae7-4308-a0a9-e3b2f44c74d2
Persistent Identifier	› 10.4121/uuid:f715aef4-6ae7-4308-a0a9-e3b2f44c74d2
Related Dataset	› The role of prenucleation clusters in surface-induced calcium phosphate crystallization - stage 3
Related Dataset	› The role of prenucleation clusters in surface-induced calcium phosphate crystallization - stage 2
Related Dataset	› The role of prenucleation clusters in surface-induced calcium phosphate crystallization - stage CaCl2 + K2HPO4
Related Dataset	› The role of prenucleation clusters in surface-induced calcium phosphate crystallization - stage 4
Related Dataset	› The role of prenucleation clusters in surface-induced calcium phosphate crystallization - stage 5
Related Dataset	› The role of prenucleation clusters in surface-induced calcium phosphate crystallization - stage 1
Related Article	› The role of prenucleation clusters in surface-induced calcium phosphate crystallization
Related	› http://rsb.info.nih.gov/ij/plugins/DM3_Reader.html
Related	› http://rsbweb.nih.gov/ij/
Related	› http://repository.tue.nl/693773
Metadata	› XML
Source	4TU.Centre for Research Data

Figure 27: Screenshot of a full record of a 4TU dataset in NARCIS.

Figure 28 below shows a PID Graph for that dataset. There are several links to other versions of the dataset through DOIs, but also to a related article that is not in the 4TU archive itself but in the repository of the Technical University of Eindhoven (TU/e). NARCIS can retrieve the title and content type (in this case article) and show it all in context.

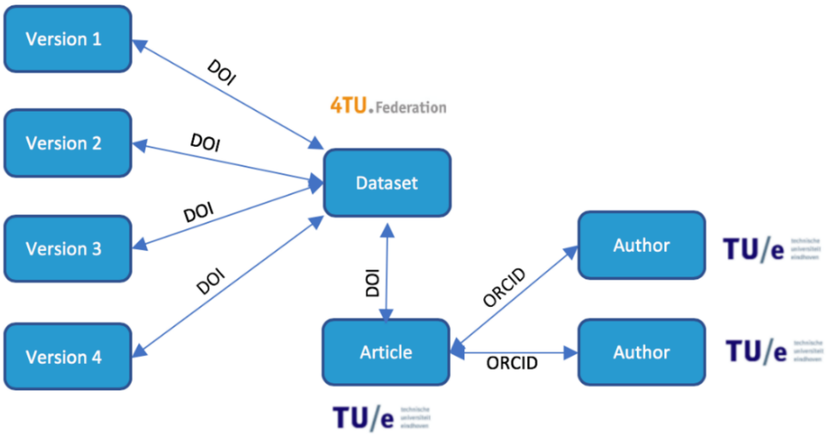


Figure 28: Graph for the 4TU dataset in NARCIS with links to different versions and an article.

5.3.5.4 Use Case: DANS EASY Archive

In two of the above use cases the EASY archive played an integral role. However, we felt the need to improve EASY to support the practice of linking to other objects. In the Humanities and Social Sciences community this awareness still needs to grow.

Currently, we are working on a new deposit interface for EASY where PIDs have a more prominent role. The interface will support references to other work or data through PIDs. Once the relations have been specified, the titles of these related resources are automatically retrieved from doi.org. Organization and person IDs will also be supported in the deposit form. The new EASY interface is scheduled for release in the beginning of 2019.

5.4 Future work: Integrating new PID types

For future FREYA tasks for this Work Package, DANS will explore the possibility of integrating organization identifiers into EASY. Currently, it seems that getting an ISNI ID for an organization is rather difficult and the only way to obtain one is through Ringgold.

In addition to motivating the technical implementation presented in the above chapters, the FREYA project also contributes to increasing awareness of PIDs and the benefits of citing, referring, resolving, and unique identification among the participating organizations in NARCIS. Via the DANS use cases we aim to encourage the use of PIDs at a national level. The use cases clearly show the potential of article-data linking and the need to establish a practice of using PIDs for these relations. Only PIDs can provide article-data linking that is long-term and persistent. We expect that the implemented use cases will lead to more extensive use of PIDs for referring to other work.

The foundation for further additions to the NARCIS PID Graph is already being laid. This will be further developed as part of the next FREYA tasks regarding emerging PID types. DANS intends to integrate emerging PID types into NARCIS. Figure 29 below is a visualization of the future NARCIS PID Graph.

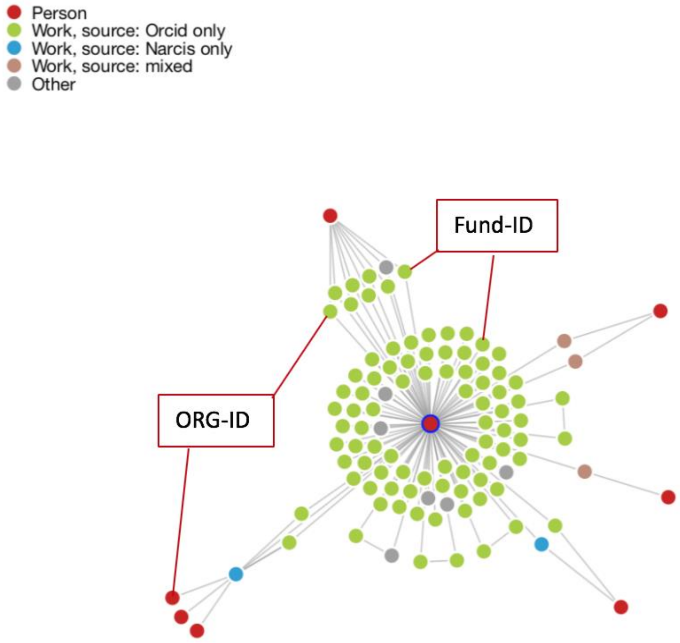


Figure 29: Emerging PID types such as Fund-ID and ORG-ID will be added to the PID Graph.

6 European Molecular Biology Laboratory (EMBL–EBI)

6.1 Background and introduction to services

The European BioInformatics Institute in Cambridgeshire UK is one of six sites that make up the Life Sciences research institution known as EMBL - the European Molecular Biology Laboratory[30]. The EBI is "home to big data in biology" and a center of expertise for the organization and analysis of this data. As a partner on the FREYA project it serves to represent life scientists, their research practices and how this influences the uptake and utilization of persistent identifiers and services.

The credit system for researchers focuses on publication of research articles that describe the major findings of research contributed to by the researcher. Publications are ordinarily multi-author articles that appear within scientific journals after being endorsed via peer review and after one or more rounds of revision. Scholarly monographs do not feature in research outputs; preprints (research articles that are posted or published ahead of time-consuming peer review) reflect a tiny percentage of published research in the Life Sciences (1.3%), but are increasing in number and recognition as a more nascent proxy for research outputs. Literature databases such as Europe PMC and PMC are the go-to repositories for indexing of published biomedical research.

When research in the Life Sciences results in the generation of large datasets (e.g. comprising genetic/protein sequences, molecular structures, atlases, or molecular archives), it is common practice for these datasets to be deposited in data repositories, many of which are freely and publicly accessible. Indeed, EMBL-EBI is publicly-funded to provide freely available data and tools (Cook et al. (2018)). There is growing recognition by stakeholders of the importance of data-sharing in the Life Sciences, with data availability statements in Life Science publications becoming more commonplace (Levchenko (2018)). The Open Access publisher PLOS, for example, has a comprehensive open data policy that gives a good indication of extent of data sharing by listing acceptable and unacceptable data-sharing methods[31].

The governance of databases in the Life Sciences is complex: each database has its own governance policies. The role of EMBL-EBI is to offer a geographic location for many publicly-accessible databases, but it does not play a central governing role. ELIXIR[32] is an umbrella organization that coordinates data provision throughout Europe. The ELIXIR hub coordinates the work of its 27 nodes, of which the EMBL-EBI is one. It provides a mechanism for coordination but not governance. Notably, many repositories based at EMBL-EBI are also considered to be ELIXIR core data resources: "of fundamental importance to the wider life-science community and the long-term preservation of biological data"[33].

Persistent identifiers are most prevalent in the Life Sciences: DOIs are used for journal articles but sparingly for data where the use of accession numbers (compact identifiers) are the norm; the use of ORCID iDs is on the increase.

Given the community norms described above, the vehicles via which the EBI is committed to respond to the mission of FREYA and its predecessor projects, include:

- identifiers.org[34], a redirection service for compact identifiers (which include accession numbers). This is a service for those implementing persistent, machine-resolvable citation of (non DOI) research data. It can now resolve any given identifier from over 600 source databases to its original source on the Web, using a common registry of prefix-based redirection rules.

[30] EMBL: https://www.embl.org/
[31] Data availability statements in Life Sciences: https://journals.plos.org/plosone/s/data-availability
[32] ELIXIR: https://www.elixir-europe.org/
[33] ELIXIR Core Data Resources: https://www.elixir-europe.org/platforms/data/core-data-resources
[34] Identifiers.org: http://identifiers.org/

- Europe PMC[35], a public repository for biomedical research literature. Built in partnership with PMC USA, Europe PMC hosts 34 million literary records (including published journal articles - full-text and/or abstracts, theses, preprints); most importantly, Europe PMC adds value to the core content by providing extensive data links. The repository and its knowledge base is supported by 29 international science funders and recognized as an ELIXIR core data resource.
- Biostudies[36] serves to hold descriptions of data underlying biological studies and provides a means to collate and find all of the data behind a research article, from resources used (e.g. sequences deposited in public nucleotide databases) to supplementary files specific to the study (these can be deposited and stored directly in the Biostudies database).

6.2 A literature-centric local PID Graph

The graphs below (Figures 30 & 31) illustrate the entities that are currently linked to the literary content in Europe PMC.

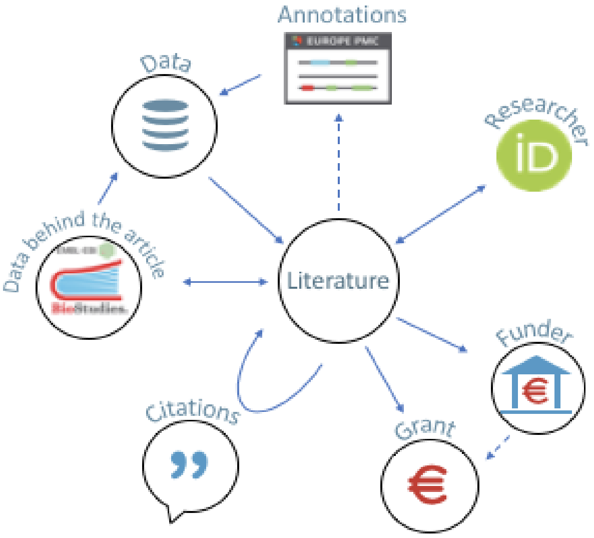


Figure 30: A literature-centric local PID Graph at Europe PMC. Europe PMC offers services that connect indexed biomedical literature to other entities (nodes, represented by circles). These include data, researchers, funders, grants, and related literature (via citations). Solid arrows are connections that rely on PIDs, dotted lines/arrows denote an absence of PIDs or PID services.

Literature-data links are best established within Europe PMC. The biomedical literature is rich in data and Europe PMC provides services that directly link the data mentioned within its literary content to its source via actionable identifiers. This is done in a number of ways. Biological terms are identified via text mining and linked to related data records in public resources. The biological terms are highlighted within the content (then known as annotations) and include gene/protein names, organisms, diseases, chemicals, Gene Ontology terms, database accession numbers, phosphorylation events, gene functions, gene-disease associations, as well as protein-protein interactions. Text mining efforts are added to by external partners including database curators, text mining groups and developers who have access to full-text literary content (that is licensed for reuse, i.e. open content), as well as all Europe PMC's text-mined annotations via APIs and are able to add their findings back as "external links".

To make it easier to access all primary data associated with a study, Europe PMC has integrated with the BioStudies database. As mentioned earlier, a BioStudies record serves as a data container, wherein the supplemental data and linked data residing in public repositories can be listed. The large majority of

[35] Europe PMC: http://europepmc.org/
[36] Biostudies: https://www.ebi.ac.uk/biostudies

Biostudies records have been created retrospectively for data-containing, full-text Europe PMC content. After an article is ingested in Europe PMC, linked data is identified by text-mined accession numbers for over 20 major data resources (Kafkas et al. (2013)) in the Life Sciences, including ENA, PDBe, and UniProt. These and any supplemental files are aggregated into a Biostudies record, assigned an accession number (with an S-EPMC prefix), which is then linked to the article identifier. The ultimate goal for linking data with literature is for data links to be set up as they are generated by a researcher so that the BioStudies identifier can be included in the research article before publication. There are currently relatively few of these records made for pre-publication (they are identifiable by the prefix S-BSST).

Literature-data links can be seen in the top left quadrant of Figure 30 above: data links are represented by “annotations” established through text mining, cross-references from data deposited in public repositories, and through integration with BioStudies.

There are various other integrations at Europe PMC concerning literature links. These services and connections are openly available via RESTful APIs[37]:

- Literature-researcher links: Integration with ORCID enables biomedical researchers to link their literary works in Europe PMC to their ORCID profiles. To date, over 6 million literary works in Europe PMC have been claimed to ORCID records.
- Literature-literature links = Citation network: The Citation network in Europe PMC[38] is generated from open citations available via Crossref services complemented with citation data from PMC full-text articles. Citations are matched to Europe PMC records in about 80% of the cases. Currently this results in approximately 300 million citation links in the network, including 13 million citing and 19 million cited articles. The network includes Life Science preprints and patents, although their contribution is minor at present.
- Literature-grant/funder links: Europe PMC provides access to funding information (grants), where available, for literary content. In addition to data provided by PubMed (USA), Europe PMC links to a database for grant information[39] from its 29 funders[40].

Figure 31 below reveals the extent of the knowledge base offered by Europe PMC, which is far greater than the literary content it contains. The numbers of unique identifiers (PMC IDs, data accession numbers, text-mined terms, ORCID iDs, grant IDs) are interesting, but more impressive are the numbers of links between these entities and their PIDs. Literature-data links comprise the largest set of connections within Europe PMC.

[37] Europe PMC developer resources: https://europepmc.org/developers
[38] Europe PMC citations network: https://europepmc.org/Help#citationsnetwork
[39] Grant finder database: https://europepmc.org/grantfinder
[40] Europe PMC funders: https://europepmc.org/Funders/

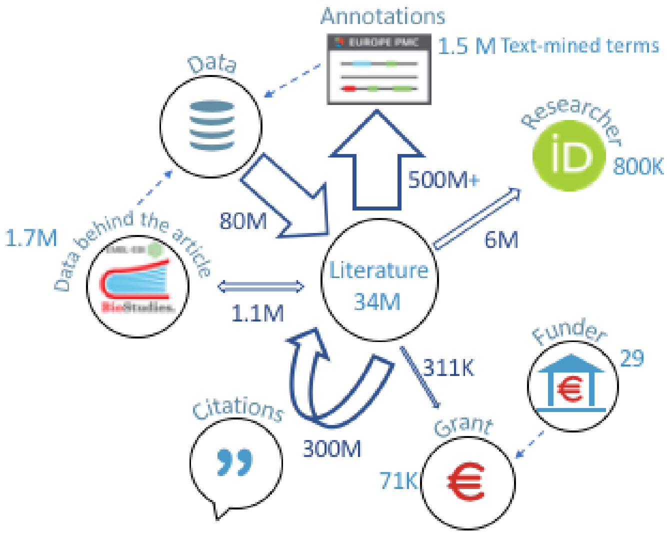


Figure 31: The local PID Graph for Europe PMC: numbers of unique entities are represented in light blue; numbers of links are represented in dark blue. Note that the numbers are correct for Q3 2018. The solid arrows are weighted anecdotally to reflect the link magnitude (note these are not directly proportional); links where numbers are not provided are represented by dashed arrows.

6.3 Current implementations

6.3.1 Adoption of schema.org by Life Science repositories

An earlier report by FREYA partners (FREYA (2018a)) pointed out that repositories in the Life Sciences provide access to their metadata largely via API downloads and therefore there is little consistency in the information that is exposed. The use of schema.org markup to encode metadata on repository landing pages would provide consistently structured information - that in turn would allow search engine optimization, for example. Currently, there are some early adopters of schema.org among repositories at the EBI. Bioschemas[41] (an extension of schema.org seeking to better define types of information specific to the Life Sciences) is less widely implemented at EBI but seen to be gaining traction. FREYA partners are conducting outreach to repositories housed at EBI to promote schema.org and encouraging upgrades. The application of digital badges to recognize uptake is being investigated.

6.3.2 ORCID claiming

An ORCID article-claiming service allows end users of Europe PMC to select and export their articles to ORCID so as to reflect their scholarly work on their profiles. Although this was made available in August 2013, constant outreach is required to promote uptake by Life Science researchers. In 2018, Europe PMC began including preprints from a number of Life Science preprint servers in its literary content alongside traditional journal publications and have been encouraging authors to claim these to their ORCID profiles. To date in excess of 14.000 preprints in Europe PMC have been linked to an ORCID iD. Given that preprints are posted or published prior to formal peer review, they are not considered “journal articles” in the traditional sense and the metadata associated with preprints is required to reflect this nuance. As a result of Europe PMC’s efforts, ORCID has agreed to include in the near future a category for “preprints” within its list of “work” types.

Like for literature-researcher links, a web service was developed during the THOR project for researchers to claim datasets to their ORCID profile, that they have helped produce/curate. This web service currently operates at EMBL-EBI and is for claiming datasets that are specifically located in repositories housed on

[41] Bioschemas: http://bioschemas.org/

site[42]. Currently five data repositories have integrated the RESTful web service which is hosted by the EBI Search engine. FREYA partners are conducting a program of outreach to promote integration by additional repositories at EMBL-EBI, as well as to encourage researchers to make use of the retroactive data claiming service.

6.3.3 Data discoverability via BioStudies

The ideal is to have data links set up as they are generated by a researcher and included in the research article before publication. Europe PMC and BioStudies are working with publishers to develop services that would make data citation easier for researchers to incorporate in their works ahead of publication.

6.4 Upcoming plans for integrations

6.4.1 Grants

Europe PMC is keeping a watchful eye on community pilots for integration of grants. These include pilots for a global ID system for grants (Kiley et al. (2018)) and for tracking research outputs of awarded grants through ORCID records[43].

Currently, Europe PMC offers a trusted resource for grant information in the form of its GRant Information SysTem (GRIST). The GRIST database is loaded with grant information from its 29 funders which are mainly European funding agencies. GRIST has established a definitive set of grant data, including funders, grant dates, name(s) of principal investigators, institutions and ORCID iD. GRIST also captures lay and technical abstracts in different languages. GRIST enables the literature-grant links represented in Figure 31 above (lower right quadrant). Grant information can be queried by end users of Europe PMC via the “grant finder” tool or by developers via the dedicated RESTful API[44]. When community pilots are complete, a goal would be to extend connections between literature and grants to include grant information from agencies beyond Europe PMC funders.

6.4.2 Research organizations

Europe PMC will seek to integrate identifiers for research organizations when pilots such as those conducted by the ROR community are complete.

[42] ORCID Claiming at EBI: https://www.ebi.ac.uk/ebisearch/documentation.ebi#af-ORCID-claiming
[43] ORCID funder working group: https://orcid.org/about/community/working-group/funders
[44] Grants RESTful (GRIST) API: http://europepmc.org/GristAPI

7 PANGAEA

7.1 Introduction

PANGAEA®[45], the Data Publisher for Earth & Environmental Science, is an Open Access publisher and library for georeferenced data from earth system research. Observational and analytical data files are archived with a description (metadata) in a relational database. The system guarantees long-term availability of its content through a commitment of the hosting institutions.

7.2 The PANGAEA PID Graph

At PANGAEA, each dataset can be identified, shared, published, and cited using a unique DOI, which is assigned to the dataset by PANGAEA during the data publication process. Hence, all datasets in PANGAEA have a DOI. To increase the (re)usability and discoverability of the published datasets, PANGAEA consistently works on including other mature persistent identifiers in the data and metadata, allowing easy access to additional information and creating persistent links between information entities related to the published datasets. Currently, PANGAEA has implemented PIDs for authors (ORCID iDs), journal articles (DOIs) and physical samples (IGSNs), which are added in the metadata whenever possible. IGSNs are to some extent still an emerging PID (see FREYA Deliverable 3.1). However, as PANGAEA has close ties to the Bremen Core Repository (BCR), which curates and stores marine sediment cores for various international ocean drilling programs (e.g. IODP), the implementation of IGSNs has taken priority at PANGAEA.

The following PID Graph (Figure 32) illustrates the linking of mature PIDs in PANGAEA.

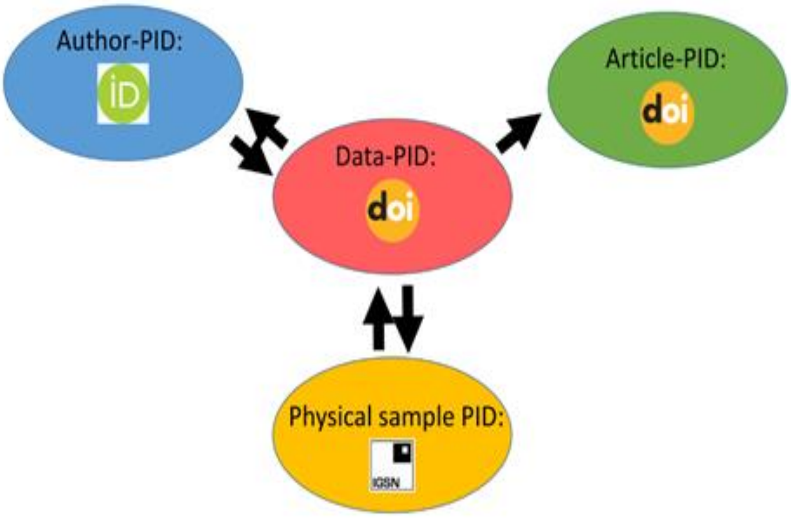


Figure 32: PID Graph depicting PIDs currently linked in PANGAEA.

PANGAEA is working on making the implemented PIDs actionable (clickable), so that they resolve to the landing page of the specific PID and associated metadata. All article DOIs and ORCID iDs are actionable. Implementation of IGSNs is ongoing and so only a subset is currently actionable.

[45] PANGAEA: https://www.pangaea.de

7.3 Current PID integrations at PANGAEA

7.3.1 Maturity level of PID integration

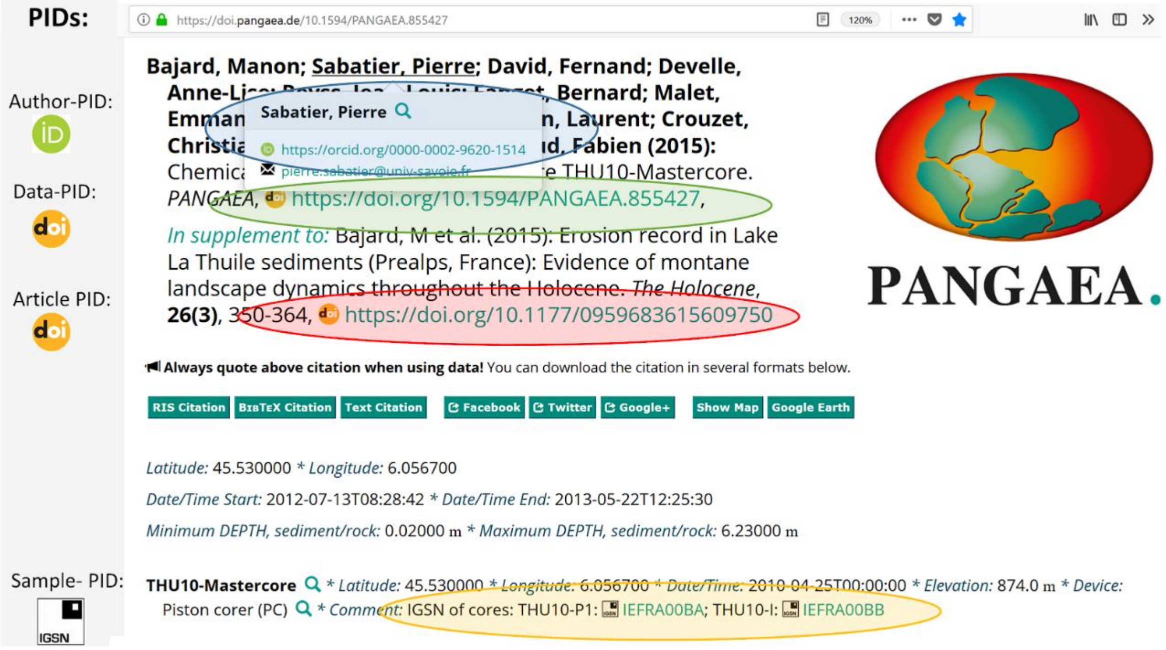
The implementation of PIDs in the metadata of a dataset depends to a large extent on the cooperation of authors in providing information on relevant PIDs. PANGAEA strongly recommends including PIDs for all submitted datasets, but there are no restrictive policies installed that would force PID integration. Integration of PIDs by PANGAEA has reached the following maturity level:

- Data PIDs: All datasets published by PANGAEA are assigned a DOI. The DOI assignment is conducted by PANGAEA during the curation and publication process.
- Article PIDs: To the extent that the data is related to a scientific publication, the PID for the publication is included in the metadata. Since authors are generally interested in generating awareness of their publication and DOIs for their scientific journal articles are already provided by the publisher, all datasets in PANGAEA, which are linked to an article, have article DOIs integrated. It should be noted that PANGAEA also handles many datasets which are not related to a specific scientific article.
- Author PIDs: PANGAEA has implemented PIDs for authors (ORCID iDs), allowing the dataset to be directly linked to the author records at ORCID. The implementation of ORCID iDs is dependent on whether the author has an ORCID iD. PANGAEA has installed initiatives promoting ORCID to authors, but this is still on a voluntary basis. Also, many of the datasets in the PANGAEA database were submitted before ORCID iDs became a widely-used PID for authors.
- Physical sample PIDs: The implementation of PIDs for physical samples in the form of an IGSN number (for further information see Deliverable 3.1) is still in its starting phase. A total number of 512 datasets have been linked to the physical sample from which they originated by including the IGSN number in the metadata.

7.3.2 Use cases

7.3.2.1 Use case: Demonstrating the usability of PIDs at PANGAEA

In the following example, we demonstrate how the implementation of PIDs in the metadata leads to an increased amount of information available for the data user. In recent years, a great deal of geological research has been conducted in Lake La Thuile, in the Northern French Prealps, to investigate the erosion patterns on geological timescales using large sediment cores. A researcher interested in getting an overview of the data available from Lake La Thuile can conduct a search in the PANGAEA database for “Lake La Thuile sediment”. This search retrieves 12 available datasets. The researcher is particularly interested in a dataset on the chemical composition of a specific sediment core and takes a closer look at the information available (Figure 33).



Bajard, Manon; Sabatier, Pierre; David, Fernand; Develle, Anne-Li **Bernard; Malet, Emman** **n, Laurent; Crouzet, Christi** **ld, Fabien (2015):** Chemic e THU10-Mastercore. *PANGAEA*, https://doi.org/10.1594/PANGAEA.855427,

Sabatier, Pierre
https://orcid.org/0000-0002-9620-1514
pierre.sabatier@univ-savoie.fr

In supplement to: Bajard, M et al. (2015): Erosion record in Lake La Thuile sediments (Prealps, France): Evidence of montane landscape dynamics throughout the Holocene. *The Holocene*, **26(3)**, 350-364, https://doi.org/10.1177/0959683615609750

Always quote above citation when using data! You can download the citation in several formats below.

RIS Citation | BIBTEX Citation | Text Citation | Facebook | Twitter | Google+ | Show Map | Google Earth

Latitude: 45.530000 * *Longitude:* 6.056700

Date/Time Start: 2012-07-13T08:28:42 * *Date/Time End:* 2013-05-22T12:25:30

Minimum DEPTH, sediment/rock: 0.02000 m * *Maximum DEPTH, sediment/rock:* 6.23000 m

THU10-Mastercore * *Latitude:* 45.530000 * *Longitude:* 6.056700 * *Date/Time:* 2010-04-25T00:00:00 * *Elevation:* 874.0 m * *Device:* Piston corer (PC) * *Comment:* IGSN of cores: THU10-P1: IEFRA00BA; THU10-I: IEFRA00BB

Figure 33: Dataset from the PANGAEA database about the chemical composition found in sediment core "THU10" from Lake La Thuile, France.

Several PIDs for information related to the dataset of interest are available through the metadata (Figure 33). The dataset itself has a DOI assigned by PANGAEA during the submission of the dataset. The data is authored by Dr. Manon Bajard, who did not provide an ORCID iD along with the data, while the co-author, Dr. Pierre Sabatier, is identifiable by an ORCID iD. The data is identified as a supplement to a research article from the scientific journal "The Holocene", which is identifiable through the article DOI provided. Finally, the IGSN sample number of the physical sediment core that the data originated from is also available.

The following additional information can also be extracted by the data user by following the PIDs available in the metadata:

- ORCID iD: In PANGAEA, the ORCID iD is actionable, resolving to the landing page of a specific researcher linked with the ID. In this example, due to the lack of an author PID for the first author, we follow the link to the ORCID iD of the second author, Dr. Pierre Sabatier. His ORCID profile reveals information on 19 other journal articles authored or co-authored by Dr. Sabatier. All of these articles are actionable and identifiable by a DOI. Furthermore, the employer of Dr. Sabatier is revealed: Université Savoie Mont-Blanc, Chambery, Rhône-Alpes, FR. This institution is identifiable by the organization identifier (ISNI). This ID is not actionable but can be found at www.isni.org. By searching for the ORCID iD of Dr. Sabatier in the PANGAEA database it is possible to retrieve 43 other datasets authored or co-authored by this researcher.
- DOI for articles: In PANGAEA, DOIs for articles related to data within the database are actionable, resolving to the landing page of the journal article. A search for the DOI of this article in the PANGAEA database reveals the dataset and other datasets originating from the same collection.
- IGSN for samples: The IGSNs for these particular sediment core samples are actionable in PANGAEA resolving to their specific IGSN landing pages. This reveals metadata including the samples' description, the geolocation from where the samples originated and the collection that the samples were part of. Searching for the IGSNs in the PANGAEA database reveals 11 other datasets originating from this specific sediment core.

7.4 Future extensions of the PANGAEA PID Graph

Using the same dataset as an example, the secondary links to ISNI and other article DOIs discovered through links to ORCID, allow for an extension of the PID Graph for this particular dataset (Figure 34).

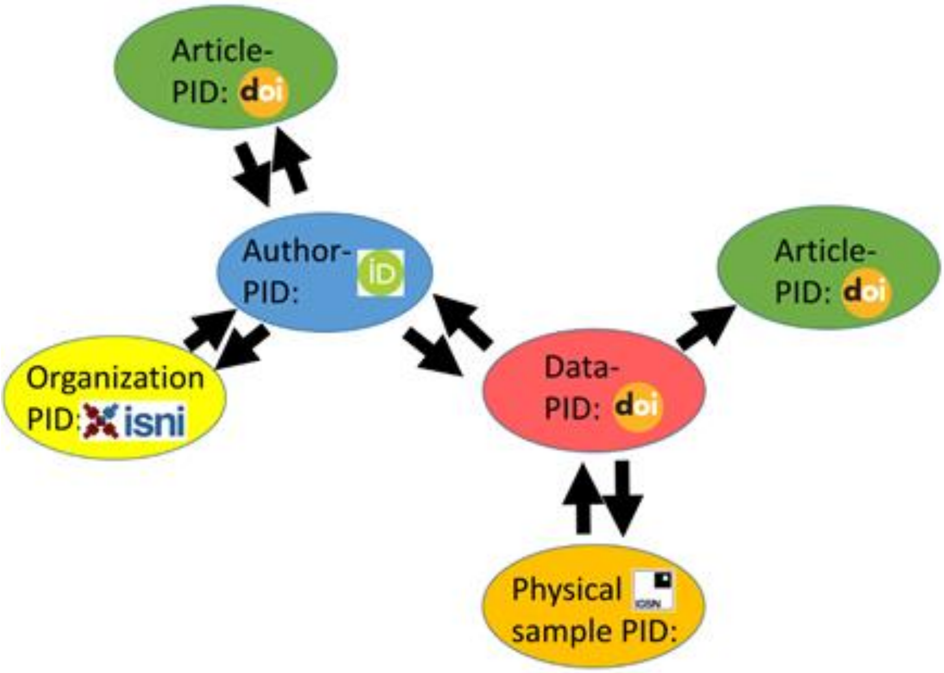


Figure 34: PID Graph specific for the example dataset regarding the chemical composition of a specific sediment core from Lake La Thuile, France.

Future work by PANGAEA will also go towards including PIDs for software, organizations and funding bodies in the metadata, so that a search starting from an IGSN number in PANGAEA can eventually return not only the data generated from scientific publications linked to the sample, but also information about who funded the research, which organizations were involved in the research and how data was generated from the sample of the core. Other PIDs (e.g. cruises and instrumentation) are also being considered, but currently have not reached a maturity level that allows for implementation.

8 Science and Technology Facilities Council (STFC)

8.1 Introduction

STFC is the Science and Technology Facilities Council[46] in the United Kingdom. Facilities Science or "lab science" is a branch of research performed by visitor scientists on large-scale instruments: synchrotron radiation and neutron sources, powerful lasers, telescopes, or supercomputers. The Facilities Science business models and research life-cycles are similar across different instruments and geographical locations, and involve extensive management of data and other information artefacts.

Apart from operating facilities, STFC is a funder for UK universities and the UK researchers that require access to large-scale research facilities across the globe, exemplified by the European Southern Observatory in Chile. STFC is also a funder of projects and of PhD studentships in the UK, and supports international PhDs and other international visitor scientists by granting them time on STFC-operated facilities. This brings a funder's perspective to the PID graphs that STFC can contribute to and benefit from.

The use case that STFC has selected to elaborate as part of its work for FREYA is focused on PhD students and the outcomes of their research. This use case represents all the aforementioned facets of the STFC operation: as an operator of large-scale facilities with multiple instruments, and also as a funder that sponsors researchers directly or through financing their projects.

The elaboration of the PhD use case requires communication with multiple stakeholders within STFC: three large-scale facilities, the library, the impacts team, and the software engineering group that develops software for the institutional repository and for interfaces with external infrastructures such as DataCite. As this presents a substantial communication overhead, a common seminar has been set up in STFC to discuss the PhD use case with information practitioners across all aforementioned stakeholders.

The PhD use case also presents a good opportunity to collaborate with other FREYA partners, in particular with the British Library who run EThOS - the UK national repository of theses. This can be a good example of how PID graphs provide a novel information infrastructure for making institutional and national information repositories interoperable, and therefore may be a good example, in terms of technology and governance, for other institutional and national repositories in Europe to follow. Another opportunity to scale this use case up is its propagation across the international network of theses repositories with a thriving community of information practitioners working for them[47].

8.2 Current STFC graph for PhD theses and related entities

The current graph is an abstraction (decomposition) of records in the institutional publication repository ePubs[48] and the Diamond bibliographic database[49]. It includes entities for theses (papers), PhDs (persons), the PhDs' host universities (organizations), facilities (PhD funders in-kind - through facility time awards - and direct funders of some PhD studentships), STFC (funder) and the Wellcome Trust (funder for one of the facilities - Diamond Light Source). This graph is represented by Figure 35.

[46] STFC: https://stfc.ukri.org/

[47] Networked Digital Library of Theses and Dissertations: http://www.ndltd.org/

[48] ePubs: https://epubs.stfc.ac.uk/index

[49] Diamond Light Source publications database: http://publications.diamond.ac.uk/pubman/searchpublicationsquick

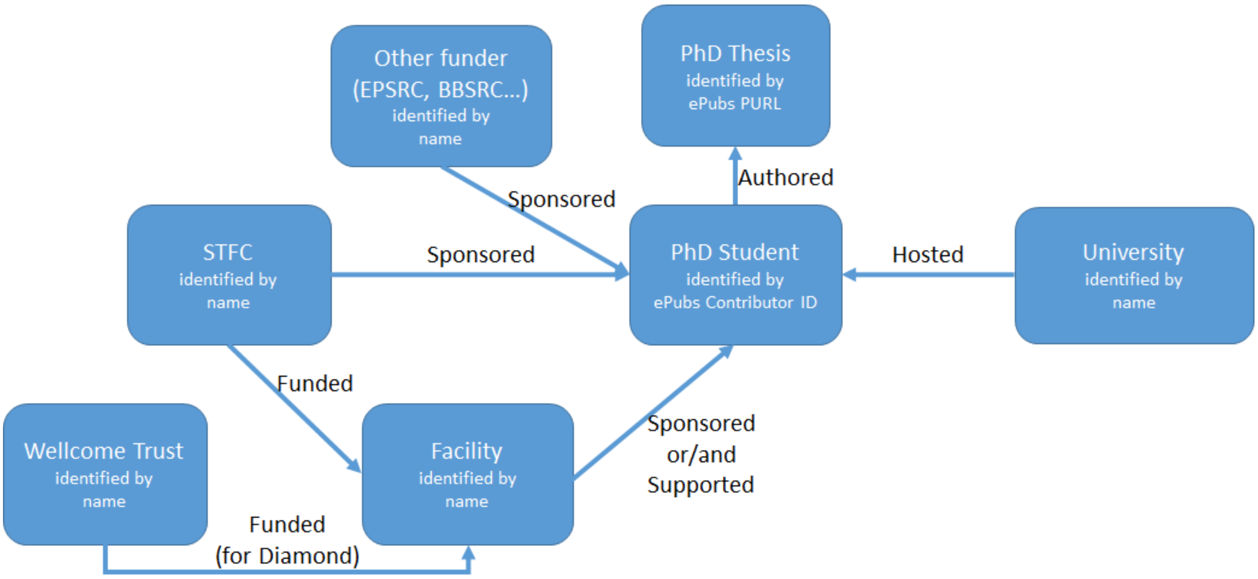


Figure 35: Current graph of entities around PhD theses supported by STFC facilities.

A specific challenge of enriching this graph with persistent identifiers is that, apart from theses and (on some occasions) PhD students, other information entities in the aforementioned metadata sources are not assigned any PIDs, and do not necessarily use terms from controlled lists or vocabularies. In many cases, the information entities, like university names, are just free text. Therefore, PIDs have to be harvested from external sources in cases like that, and then associated with the respective metadata fields in the institutional repositories.

8.3 Current PID implementations at STFC

Current PID implementations at STFC include:

- ORCID iDs for authors: when the ORCID iD for a particular author does not exist or is not known in the institutional publications repository, a Contributor ID is used, which is a specific stable ID that is minted by the STFC publications repository.
- DOIs for datasets are used in the institutional data repository. These DOIs have limited potential for their inclusion in the graph for the PhD use case, but they can be used in other contexts.
- DOIs for facility experiments (investigations) minted via the DataCite API. The landing pages for these DOIs are supported by STFC facilities. The landing pages refer to raw data collected during facility experiments and to large-scale facility instruments where experiments have been performed.
- DOIs for research papers: each publication in the institutional repository, irrespective of it having or not having a DOI associated, is also identified by a PURL minted by the STFC library. This PURL resolves to a landing page in the STFC publications repository.

8.4 Extensions of STFC graph for PhD theses and related entities

The expected evolution of the STFC graph for PhD theses and related entities is represented in Figure 36, which also indicates the role of other STFC stakeholders beyond the core FREYA team, whose participation will be required to actually deliver this advanced graph.

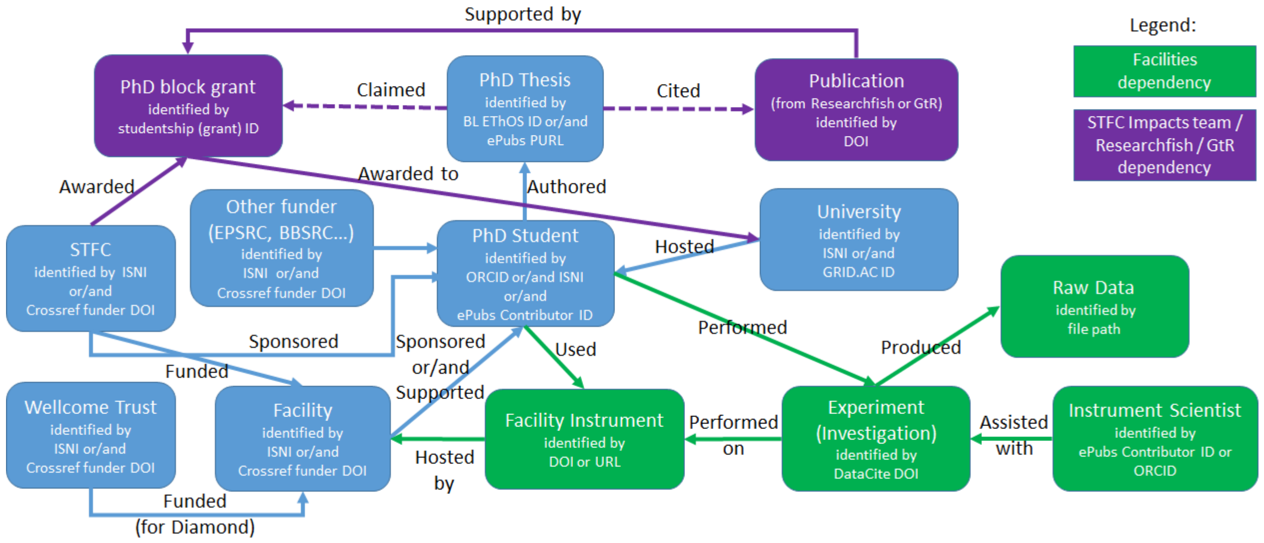

Supported by
Legend:
Facilities dependency
STFC Impacts team / Researchfish / GtR dependency
PhD block grant identified by studentship (grant) ID
Claimed
PhD Thesis identified by BL EThOS ID or/and ePubs PURL
Cited
Publication (from Researchfish or GtR) identified by DOI
Awarded
Awarded to
Authored
Other funder (EPSRC, BBSRC...) identified by ISNI or/and Crossref funder DOI
STFC identified by ISNI or/and Crossref funder DOI
PhD Student identified by ORCID or/and ISNI or/and ePubs Contributor ID
Hosted
University identified by ISNI or/and GRID.AC ID
Raw Data identified by file path
Performed
Produced
Sponsored
Sponsored or/and Supported
Used
Funded
Wellcome Trust identified by ISNI or/and Crossref funder DOI
Facility identified by ISNI or/and Crossref funder DOI
Facility Instrument identified by DOI or URL
Performed on
Experiment (Investigation) identified by DataCite DOI
Assisted with
Instrument Scientist identified by ePubs Contributor ID or ORCID
Hosted by
Funded (for Diamond)

Figure 36: Future graph of entities around PhD theses supported by STFC facilities, with more entities identified by PIDs and with new entities added.

The use case of PhD theses and related entities will be elaborated together with the British Library and their team who run the EThOS repository. The experiments of matching records between the STFC publications repository and EThOS have started. The collaborations with other stakeholders within STFC and beyond the core FREYA team that will be required are indicated in green and violet in Figure 38.

STFC foresees a need for assigning multiple PIDs for the same information entity. Examples of this are ISNIs and GRID IDs for organizations or ISNIs and ORCID iDs for persons. The reason for this is twofold. Firstly, some PIDs like grant numbers may be favored by policy makers, but they may not follow the best practice of having a PID resolved into a quality landing page; they are just identifiers accompanied by descriptions in a centralized database which may not even be publicly available or only partially published. Therefore, other identifiers with better publishing practices can complement such PIDs. Another reason why the association of multiple identifiers with the same entity may be beneficial is that this allows crosswalks between different identifiers. This can support future services that will allow requesting PID “synonyms” across PID providers or services that build bigger graphs from diverse metadata sources that use various systems for identification of entities of the same type.

9 Conclusions

This first report on the integration of the FREYA PID Graph into disciplinary contexts underlines how PIDs and the PID Graph can be used and applied in different communities despite the diversity of their settings and respective challenges. The presented pilot applications take their use cases from a variety of communities, ranging from the Life Sciences, Humanities and Social Sciences to High-Energy Physics. This means that services that are part of the pilot applications face different challenges (e.g. scalability or different requirements for metadata granularity) and have distinct ambitions and approaches. Such differences are most evident when comparing the graphs presented in each partner's chapter or the setup and level of advancement of these pilot applications - the underlying service infrastructure and operations vary significantly among the pilot applications; variation can also be observed in the use of slightly different terminology, e.g. services vs. use cases. We anticipate that some of these inconsistences (most importantly the dissimilarity among graphs) will become significantly smaller during the lifetime of the project through our work on improving and upgrading these community PID graphs.

These observations are particularly noteworthy when discussing the global Open Science and scholarly communication communities which are even more diverse. Having demonstrated that we are able to prototype and implement mature and even emerging PID types in these diverse settings underlines the potential of PIDs in the context of EOSC and other scalable infrastructure projects. In short, it can be argued that developing meaningful connections through persistent identifiers to advance the building of the PID Graph can be accomplished regardless of disciplinary or community complexities.

Studying these implementations by the FREYA partners, the PID Graph is being used as a facilitator to complete the scholarly record to the benefit of science. The integrations build more connections, and those connections are made accessible and visible to users. By doing so, such information can be corrected and used in the research and publishing process. The pilot applications are a means for introducing the PID Graph to users. This means, first of all, that research resources with PIDs are more visible to them (because PIDs, like DOIs, enable findability) and, secondly, that users see connections between resources, e.g. article-data links. Building on the work presented in Deliverable 3.1, the integrations presented here underline the maturity of person-article-data linking, which is part of each of the pilot applications (in various forms). Similarly, PIDs for software/code is also a key topic that several partners have worked on, but the level of maturity is not equal to that of data PIDs, for example. The use of DOIs for software and code has definitely reached the mainstream in Open Science. This is encouraging as it helps making the scholarly record more complete and contributes to enabling research communities to practice reliable Open Science.

Looking at the work done in the pilot applications, something that is common amongst a number of them is that they are starting to build and use PIDs early in the research workflow. Thanks to the different pilot applications this can be done in various ways: e.g. by using PIDs for research outputs and linking them to grants. This can also be linked to data, code and researchers involved, so that the whole research lifecycle is covered; tracking impact and reporting to funders can then become a straightforward process.

The impact from the work carried out by this Work Package can already be seen through the newly introduced Google Dataset Search[50]. One important task many of the FREYA partners were able to tackle during the first year of the project was integrating schema.org into their various systems[51]. Structured markup allows resources hosted within these repositories to be discoverable by Dataset search. This was a crucial task for laying a solid foundation for the PID Graph and it can be considered a milestone in the context of the FAIR principles (where F = findable).

Regarding the impact of the implementation of the PID Graph on researchers in Europe, it is not expected that these first steps presented here will have an effect. The main objective behind the PID integrations by

[50] Google Dataset Search: https://toolbox.google.com/datasetsearch

[51] Schema.org integrations in FREYA: https://github.com/datacite/freya/blob/master/schema_org.csv

the disciplinary partners was to introduce missing links between entities and services. Researchers, i.e. users of these services, will be able to take advantage of the fact that their published output is more complete in that they might link to more resources elsewhere in a persistent way. We stipulate that even though such steps will not greatly impact researchers' work, they will smooth the transition towards improved Open Science services and are crucial for driving adoption of Open Science in the research communities. As this work continues, reporting on science progress to institutions, funding organizations or policy makers will become much easier for researchers.

Finally, an important next step for the PID Graph is to scale to EOSC, as already mentioned in previous chapters. With its open and participative architecture, the PID Graph is exploitable by any infrastructure and service. Taking this step will help ramp up Europe's Open Science practices.

10 Challenges and future work

In this deliverable, we presented considerations and implementations stemming from the first year of work carried out by the pilot applications in FREYA. While the results show a promising exploitation of PIDs, this is only the starting point. First integrations have been starting slower than expected; the potential use of PIDs enables a new thinking of sharing, preserving or publishing workflows and, in some cases, reassessment of workflows to exploit PIDs in the best way possible was necessary.

We expect that many more opportunities will unfold over the course of the next year, in particular when considering the new and emerging PID types (see the forthcoming Deliverable 3.2). The pilot applications' work with these emerging PIDs (visualized in the "future PID graphs" sections within the pilot application chapters in this report) underline that emerging or new PID types whose development is now progressing considerably, e.g. for organizations or instruments, will close considerable gaps.

A key challenge will be to assess the impact of features or changes applied. Even more so, it might take more time to notice a significant change in behavior, as we are not only discussing a technical change, but an expected cultural change which takes time. The great benefit of developing pilot applications as part of FREYA is that we are able to be the first implementers of new concepts into the workflows of our respective communities. However, continuing these integrations and observing adoption from the communities does not happen quickly.

Hence, for FREYA it will be important to collaboratively support the cultural change along with Work Package 5 (Engagement) and to prepare key performance indicators for better observation and evaluation of the work conducted. The first results have shown that the pilot applications are well prepared to be the testbeds for new and emerging PID types in the next years.

Bibliography

Anderson C. The End of Theory: The Data Deluge Makes the Scientific Method Obsolete. Wired. 2008. URL: https://www.wired.com/2008/06/pb-theory/

Bakeer M. 1,500 scientists lift the lid on reproducibility. Nature. 2016; 533: 452–454. doi:10.1038/533452a

Borgman C. L. The conundrum of sharing research data. Journal of the American Society for Information Science and Technology. 2012. doi:https://doi10.1002/asi.22634

Cook C. E. et al. The European Bioinformatics Institute in 2017: data coordination and integration. Nucleic Acids Research. 2018; 46(D1): D21–D29. doi:10.1093/nar/gkx1154

Drucker J. Intro to Digital Humanities: Introduction. UCLA Center for Digital Humanities. 2013. URL: http://dh101.humanities.ucla.edu/

Dzogang F. et al. Discovering Periodic Patterns in Historical News. PLoS ONE. 2016; 11(11): e0165736. doi:10.1371/journal.pone.0165736

European Commission. EOS Declaration. 2017. URL: https://ec.europa.eu/research/openscience/pdf/eosc_declaration.pdf

FORCE 11. The FAIR Data Principles. 2016. URL: https://www.force11.org/group/fairgroup/fairprinciples

Foster I. et al. Big Data and Social Science: A Practical Guide to Methods and Tools. Chapman and Hall/CRC Press. 2016. ISBN 9781498751407

Frees E. W. Longitudinal and Panel Data: Analysis and Applications in the Social Sciences. Cambridge University Press. 2004. ISBN 978-0521535380

FREYA. Deliverable 2.1: PID Resolution Services Best Practices. 2018a. doi:10.5281/zenodo.1324300

FREYA. Deliverable 3.1: Survey of Current PID Services Landscape. 2018b. doi:10.5281/zenodo.1324296

Hey T. et al. The Fourth Paradigm: Data-Intensive Scientific Discovery. Microsoft Research. 2009. ISBN 978-0-9825442-0-4

Kafkas Ş. et al. Database Citation in Full Text Biomedical Articles. PLoS ONE. 2013; 8(5): e63184. doi:10.1371/journal.pone.0063184

Kiley R. et al. Wellcome explains the benefits of developing an open and global grant identifier. CrossRef Blog. 2018-02-16. URL: https://www.crossref.org/blog/wellcome-explains-the-benefits-of-developing-an-open-and-global-grant-identifier/

Levchenko M. Mapping out the path to data. Data availability statements in biomedical literature. Europe PMC Blog. 2018-11-08. URL: http://blog.europepmc.org/2018/11/mapping-out-path-to-data.html

Maass W. et al. Big Data and Theory. In: Schintler L., McNeely C. (eds) Encyclopedia of Big Data. Springer, Cham. 2017.

Miguel E. et al. Promoting Transparency in Social Science Research. Science. 2014; 343 (6166). doi:10.1126/science.1245317

Molloy J. C. The Open Knowledge Foundation: Open Data Means Better Science. PLoS Biology. 2011; 9(12), e1001195. doi:10.1371/journal.pbio.1001195

Murray, S. The LSST and big data science. Astronomy Magazine. 2017 URL: http://www.astronomy.com/news/2017/12/the-lsst-and-big-data-science

Nosek B. A. et al. Promoting an open research culture. Science. 2015; 348 (6242) doi: 10.1126/science.aab2374

Open Science Collaboration. Estimating the reproducibility of psychological science. Science. 2015; 349 (6251). doi:10.1126/science.aac4716

Pampel H., Dallmeier-Tiessen S. Open Research Data. From Vision to Practice. In: Bartling S., Friesike S. (eds) Opening Science. Springer, Cham. 2014. doi:10.1007/978-3-319-00026-8_14

Shiers, J. et al. CERN Services for Long Term Data Preservation. CERN IT Note. 2016. URL: http://cds.cern.ch/record/2195937

Sitek D., Bertelmann R. Open Access: A State of the Art. In: Bartling S., Friesike S. (eds) Opening Science. Springer, Cham. 2014. doi:10.1007/978-3-319-00026-8_9

Tripathee A. et al. Jet substructure studies with CMS open data. Physical Review D. 2017; 96 (074003). doi:10.1103/PhysRevD.96.074003

Wallis, S. Annotation, Retrieval and Experimentation. In: Annotating Variation and Change. University of Helsinki. 2007. URL: http://www.helsinki.fi/varieng/series/volumes/01/wallis/

Wilkinson M. D. et al. The FAIR Guiding Principles for scientific data management and stewardship. Nature Scientific Data. 2016; 3 (160018). doi:10.1038/sdata.2016.18