



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

D4.6 Emerging and New PID Graph Resource Types in Disciplinary Contexts

Deliverable type Report
Dissemination level Public
Due date 30 November 2020
Authors Christine Ferguson (EMBL-EBI, orcid.org/0000-0002-9317-6819)
Arthur Thouvenin (EMBL-EBI, orcid.org/0000-0002-6004-756X)
Robin Dasler (DataCite, orcid.org/0000-0002-4695-7874)
Chris Baars (DANS, orcid.org/0000-0002-5228-1970)
Artemis Lavasa (CERN, orcid.org/0000-0001-5633-2459)
Stephanie van de Sandt (CERN, orcid.org/0000-0002-9576-1974)
Frances Madden (British Library, orcid.org/0000-0002-5432-6116)
Vasily Bunakov (STFC, orcid.org/0000-0003-3467-5690),
Natasha Simons (ARDC, orcid.org/0000-0003-0635-1998)
Tina Dohna (PANGAEA, orcid.org/0000-0002-5948-0980)
Abstract This deliverable focuses on the integrations of emerging PID resource types made by the disciplinary partners in FREYA, including organization IDs, grant and funder IDs. It summarizes lessons learned of use to communities wishing to undertake similar implementations. A status update is provided for PID resource types identified as emerging or immature at the outset of the project that have been moved forward since then.
Status Submitted to EC 27 November 2020

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



FREYA project summary

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

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

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

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

Disclaimer

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

Copyright Notice

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

Executive summary

This deliverable completes the task of integrating emerging PID resource types into disciplinary services which have been developed or enhanced through the FREYA project. It builds on work from previous Work Package 3 and 4 deliverables and provides lessons learned that occurred through working with emerging PIDs.

Picking up on the PID landscaping exercise conducted by partners at the outset of the project, it focuses on PID resource types identified as emerging or immature in 2018. The report begins with an update on the status of many of these resource types: first, we summarize those driven forward by FREYA partners who have developed prototypes for implementation of ROR IDs from what was at the time a “minimal viable” Research Organization Repository, grant DOIs that form part of the fledgling Global Grant Identifier System, and research instrument identifiers that form part of the “Sensor Information System infrastructure” for earth and environmental scientists. Next, the report summarises progress made within partner communities (outside FREYA) on a variety of other emerging PID resources ranging from data management plans to samples and facilities to give a more complete picture of the current status of emerging PIDs.

The main part of this report centers on further integrations by the disciplinary partners to their databases and workflows, that build on existing prototypes for PID Graph resource types. A summary of work is offered from six disciplinary partners who outline their specific system requirements for organization IDs (ROR, GRID and ISNI IDs), grant and funder IDs, noting the specifications and integrations that served to address these requirements. While the focus is very much on emerging PIDs, some updates have been included on mature PID integration for work which had been started in previous deliverables. Partners achieved integrations of identifiers for emerging PID Graph resources into existing database records “retrospectively” or provisions were made for these identifiers to be incorporated into any newly added records.

Importantly, we show that emerging PIDs can be implemented across a range of disciplines and can address a variety of use cases. By embracing nascent infrastructures and uncovering technical weaknesses, FREYA partners have helped to grow emerging PID infrastructures.

Contents

1	Introduction.....	5
2	Revisiting the status of new and emerging PID resources	6
2.1	Prototypes developed in Work Package 3.....	6
2.2	New and emerging PID resource types that have moved forward outside of FREYA project	7
3	Disciplinary integrations	11
3.1	EMBL-EBI.....	12
3.2	DANS.....	18
3.3	British Library.....	22
3.4	CERN	26
3.5	PANGAEA.....	31
3.6	STFC	38
4	Lessons learned and concluding thoughts	43

1 Introduction

At the outset of the FREYA project, a landscaping exercise was undertaken to document the PID landscape that was being contemplated by each partner and the community they represent. The report, submitted as Deliverable 3.1¹ in 2018, served to benchmark the range of research resources that had or required persistent identifier solutions at the time and commented on the relative maturity of PID infrastructures that were in operation. A maturity matrix was drawn up that reflected FREYA partners' focus and their view of PID infrastructures available to their community for each research resource (Table 1, Deliverable 3.1), rather than a comprehensive commentary on every PID resource available globally.

Infrastructures for publication identifiers, data identifiers and researcher identifiers were deemed to be mature, in that the PIDs are indexed by multiple databases and workflows, bespoke workflows and search engines had been in operation for a number of years (decades in some cases) or the infrastructures were the focus of predecessor projects to FREYA, ODIN² and THOR³. All other PID infrastructures were assigned *emerging* or *immature* status. It is research resources in these categories that we focused on for the remaining two deliverables in Work Package 3. Deliverable 3.2⁴ involved mapping the gaps in the PID landscape with user stories collected by partners followed by feasibility studies for prototype implementation. Deliverable 3.3 culminated in demonstrators of prototyping implementations conducted by the FREYA partners of new PID types and new PID services. The building work in Work Package 3 ("building what we don't have") was taken up in Work Package 4 ("incorporating it")⁵. In D4.4 "Organizational IDs in practice", a range of organization IDs were discussed with special focus on the ROR ID as a community-led initiative with open infrastructure and data that is well suited for use in an open science environment⁶. Likewise, the **aim of the current deliverable** is to report on our further integrations of some of these *emerging* and *new* PID graph resource types in disciplinary systems, databases and workflows.

Chapter 2 of this report summarizes the *new* and *emerging* PID Graph resource types that were the focus of Work Package 3. Section 2.1 elaborates on PID Graph resource types for which prototypes were worked on by FREYA partners; section 2.2 provides a summary of work on PID Graph resource types that has moved forward largely outside of the FREYA project, as well as those that are currently too immature to gain traction from the research community.

Chapter 3 consists of contributions from FREYA partners who have built further implementations to include new and emerging PIDs described above that meet specific needs of their communities. FREYA partners with the following disciplines report on their progress: EMBL-EBI (representing Life Sciences), DANS (representing Social Sciences), the British Library (representing Humanities and Social Sciences), CERN (representing High-Energy Physics), PANGAEA (representing Earth and Environmental Sciences) and STFC (representing Facilities-based Science).

The final chapter discusses lessons learned by each partner and a discussion of the value for the community at large of implementing new and emerging PIDs types.

¹ FREYA Deliverable 3.1: <https://doi.org/10.5281/zenodo.3554255>

² ODIN: <https://odin-project.eu/>

³ THOR: <https://project-thor.eu/>

⁴ FREYA Deliverable 3.2: <https://zenodo.org/record/2649230#.X5MwCUJKjUI>

⁵ Quotations represent the broad aims of FREYA's Work Packages; taken from Table 1, Deliverable 3.2

⁶ FREYA Deliverable D4.4 <https://zenodo.org/record/3666255#.X5MwC0JKjUI>

2 Revisiting the status of new and emerging PID resources

2.1 Prototypes developed in Work Package 3

In the final phase of work package 3, FREYA partners built demonstrators for prototypes of new PID resources. These focused on PIDs for organizations, for grants, for projects and instruments specifically on research vessels and PIDs for facilities. This work followed two earlier deliverables, which identified gaps in the PID landscape and determined feasibility of prototype implementation, respectively.

A report was drawn up at the conclusion of this work in February 2020 (see the report submitted to accompany these demonstrators⁷). The report provides a useful marker of the status of those efforts which must be considered in the context of the Identifier landscape at that time. Ultimately, prototypes were realised as follows: ROR IDs for Organizations was the most advanced demonstrator built, given that the infrastructure for these organization IDs was in place and being promoted globally; global grant identifiers progressed well, but Europe PMC's work with integrating these identifiers serves as a first demonstrator for how global grant IDs might be adopted by stakeholders to unambiguously connect grants to research outputs; there was some prototyping realised for instrument IDs, with larger research infrastructures minting Digital Object Identifiers (DOIs) for instruments as early adopters. However, inclusion of additional metadata fields in the DataCite schema to satisfy the instrument identification requirements has not been completed. Prototyping identifiers for facilities turned out to be complex, in that the work considered research resources beyond the facility per se (i.e. Diamond Synchrotron and beam time awarded to researchers) and considered a range of identifiers that might be required to determine the value offered by the facility. While prototyping began in Work Package 3 and represented the vision for an Open Science Portal for STFC, this facilities demonstrator has been developed in Work Package 4.

A brief recap of the prototypes that were built by FREYA partners is offered here. The ROR repository⁸, containing **ROR IDs for Organizations** had been set up since early 2019 and the infrastructure has evolved sufficiently in the year since the ROR MVP (minimum viable product) was launched for DataCite to incorporate ROR into DataCite services. This resulted in production-level services that make ROR IDs available to every DataCite member. The ROR infrastructure has continued to grow, attracting global support to date of community advisors, integrating platforms, signatories pledging to adopt ROR IDs and funding. Of the new and emerging PID resources, these organization IDs are of interest across disciplines and constitute the most robust candidates for implementation in the current deliverable. Indeed, the previous FREYA deliverable⁹ was entirely devoted to an assessment of organization IDs and early initial implementations of ROR IDs by project partners.

Grant DOIs are the solution to the quest for a global grant identifier to replace the internal identifiers assigned by individual funders to their own awards. Funders who register with Crossref¹⁰ receive DOIs in exchange for grant record metadata that can subsequently be included in publication metadata. Europe PMC has supported the Wellcome Trust in providing landing pages for its grants and sending grant metadata packages to Crossref. As a demonstrator of how the global grant ID system could work, Europe PMC worked with the publisher PLOS to link publications to Wellcome grant records and thereby to other publications funded by the same grants - all conducted via the grant IDs and Wellcome's grant records within Europe PMC's grantfinder database¹¹. The infrastructure for global grant identifiers is still in its infancy and to date only a handful of funders from Europe and the US have registered with Crossref for

⁷ Deliverable 3.3 Prototypes of New PID Resources Insert link?

⁸ <https://ror.org/>

⁹ FREYA Deliverable 4.4 <https://zenodo.org/record/3666255#.X5Mwc0JKjUI>

¹⁰ <https://www.crossref.org/community/grants/>

¹¹ <http://europepmc.org/grantfinder>

grant DOIs. Stakeholders, such as publishers, will need to be encouraged to implement these PIDs in their workflows before researchers can be incentivized to include them in their publications.

Research instrument identifiers are part of the “Sensor Information System infrastructure” set up by the Alfred-Wegener Institute (AWI), which coordinates German polar research¹². Researchers who register the instruments they use, receive a “handle” for the instrument description that can subsequently be included in data publication metadata. PANGAEA has started including these handles for instruments in dataset metadata. This initial step makes it possible to aggregate related data (based on instrument/sensor use) using machine-to-machine communication at PANGAEA. Given their focus on research vessel instrumentation, these identifiers are naturally of most interest to the Earth and Environmental Sciences community.

2.2 New and emerging PID resource types that have moved forward outside of FREYA project

This section summarizes the progress made regarding new and emerging PID resources since the landscaping exercise reported in Deliverable 3.1 in June 2018. These PID resource types were considered too immature for prototyping by individual project partners. Solutions for many of these have been moved forward by consortia outside of FREYA and we outline below any developments that have relevance for disciplinary communities of project partners.

Data Management Plans (DMPs)

It is already technically feasible to assign a PID to a DMP, and this is especially simple to do for a DMP that has been deposited in an institutional repository that provides PIDs. However, this has normally assumed that the DMP is a static text document. During the time of the FREYA project, two RDA working groups (unrelated to FREYA) have been working on separate but related aspects of viewing the DMP as a container for all of the outputs of a specific research project. The DMP Common Standards Working Group¹³ has developed a common data model with a core set of elements for describing a DMP, and the Exposing Data Management Plans Working Group¹⁴ has been exploring use cases for exposing DMPs to human and machine actors outside of the DMP’s institutional setting. The PID Graph offers opportunities for this view of the DMP as a container by enabling connections between DMPs, datasets, funding institutions, authors, and so on. To better accommodate DMPs, version 4.4 of the DataCite Metadata Schema that will be released at the end of 2020 will add an explicit resource type of “data management plan” so that these kinds of outputs can be directly tracked.

Projects

There is sometimes confusion about the difference between a grant and a project in the context of identifiers. A grant or funding award is an agreement made at a particular time that provides funding support to the researcher for a specified research activity. A grant ID refers to this agreement. In essence, a grant is a transaction between a research funding body and a researcher or group of researchers. In contrast, a project is an activity or defined sequence of activities carried out by a researcher or group of researchers. A project identifier is a compound identifier that brings together the various entities related to the project in various ways (such as researchers involved in the project, the funding grant awarded to do the project, data and articles produced, software platforms used and so on). A single project can therefore have one, several, many, or no grants at all. These distinctions between a grant and a project make it clear that an identifier is needed for each – an identifier for the grant (transaction) and an identifier for the project (activities).

¹² <https://sensor.awi.de/>

¹³ <https://www.rd-alliance.org/groups/dmp-common-standards-wg>

¹⁴ <https://www.rd-alliance.org/groups/exposing-data-management-plans-wg>

In terms of grant identifiers, research funders are joining Crossref as members to register research grants with them, to be able to accurately track this information at the individual award level (this is explained in the section 2.1). In terms of project identifiers, while there are a number of local initiatives, the most promising solution to date is the Research Activity Identifier (RAiD) which is a compound object identifier. To recap from Deliverable 3.1, RAiD was established in 2017 and is made up of an identifier (a Handle) plus an “envelope” containing associated PIDs and metadata, that can capture the individuals, organizations, funding grants, equipment, data, publications and other research entities linked to the activities of a research project. In the context of open research, RAiD also provides a record of the context for and contributors to each output, which is vital for research integrity and reproducibility.

RAiD is managed by the Australian Research Data Commons (ARDC)¹⁵, and is in active use in seven Australian organizations at the time of writing, with a further 24 having access to the system but no live integration as yet. 5,366 live RAiDs have been minted. RAiD is integrated with ORCID records and holds associations with a range of other identifier systems, including ROR, ISNI¹⁶, and GRID¹⁷ for organizations; DOIs for datasets and articles; and Handles for instruments. The RAiD team are in discussions with partners in the USA and the Netherlands who are planning pilot RAiD projects. RAiD is in the process of being certified as an official international standard with ISO¹⁸, which is expected to be published by May 2021. A project identifier such as RAiD enables research groups and institutions to associate people, data, works and funding with a long-term effort, to track the impact of these efforts over the long-term, and to focus on the narrative, rather than a particular researcher or funding stream.

Conference PIDs

The Conference ID Working Group¹⁹, chaired by Crossref, has continued its work during the lifetime of FREYA²⁰. Crossref has developed a proposed schema update to accommodate conference IDs. In the meantime, the ConfIDent project has emerged to develop a separate metadata schema for conference IDs²¹. To support interoperability, ConfIDent has proposed a set of updates to and recommendations for the DataCite Metadata Schema.

Sample PIDs

The International Geo Sample Number (IGSN) for samples has gained wider adoption since the FREYA working group’s assessment in Deliverable 3.1. While IGSNs have been primarily used by the geological scientific community to identify geological samples, the use of these identifiers can be seen worldwide and is rapidly finding use in other disciplines. IGSN 2040²² funded by the Sloan Foundation²³, has actively fostered this disciplinary expansion and is tackling some of the most pressing issues relating to sustainability, workflows and architecture of the PID infrastructure. The need for a persistent sample identifier to track the entire sample life cycle (in the field -> in the lab -> in the sample repository -> in the data repository) persists throughout the larger scientific community from the Natural Sciences to the Humanities. The IGSN 2040 project consortium seeks to re-design and mature the existing organization and technical architecture of the IGSN to create a global, scalable, and sustainable technical and organizational infrastructure for PIDs of material samples. At the same time, publishers are encouraging the use of IGSNs and large consortia, such as the Centres of the German Helmholtz Association, are considering institution-

¹⁵ The Australian Research Data Commons (ARDC): <https://ardc.edu.au/>

¹⁶ ISNI: <http://www.isni.org/>

¹⁷ GRID: <https://grid.ac/>

¹⁸ ISO: <https://www.iso.org/home.html>

¹⁹ Working group “PIDS for conferences and Projects’: <https://www.crossref.org/working-groups/conferences-projects/>

²⁰ FREYA blog post “Towards Persistent Identification of Conferences’: <https://www.project-freya.eu/en/blogs/blogs/towards-persistent-identification-of-conferences>

²¹ ConfIDent project website: <https://projects.tib.eu/en/confident/>

²² ISSN 2040 project website: <https://www.igsn.org/igsn-2040/>

²³ The Sloan Foundation: <https://sloan.org/>

wide implementation of IGSNs in their extensive and heterogeneous collections. To support this positive trend, improvements in the structure and facility of sample registration have been implemented (e.g. batch metadata updating, API for submission, links to cruise DOIs, publication and data DOIs). More than 7 million IGSNs have been issued so far by eight allocating agents.

Instrument PIDs (DOIs, Handles)

The Research Data Alliance Working Group Persistent Identification of Instruments (PIDINST)²⁴ developed a community-driven solution for persistent identification of instruments. To ensure that developments would follow community needs and requirements, the working group solicited use cases from a diverse community of potential instrument PID users²⁵. On the bases of these use cases and with the input from the use case developers, the metadata schema was developed and iteratively adapted to fit the broad spectrum of stakeholders involved. Two PID minting options are currently available to those interested in using persistent identifiers for instruments. ePIC²⁶ offers the option to mint handles for instruments using the exact metadata schema provided by the PIDINST working group. DataCite will adapt its metadata schema in future versions to allow for a more complete mapping of the proposed instrument metadata schema to the DataCite schema. FREYA Partner STFC is planning to join HZB (Helmholtz-Zentrum Berlin für Materialien und Energie) and BODC (British Oceanographic Data Centre) as an early adopter of instrument PIDs by institutional instrument providers. PANGAEA includes handles (SENSOR.AWI) for instruments in their dataset metadata and is anticipating the adoption of the PIDINST schema and DOI instrument registration by related services from the Alfred Wegener Institute (AWI) - also mentioned in section 1.1 above. Reliant on the progress of the AWI in this regard, activities around instrument PIDs with FREYA involvement have focused on connecting with other communities working on controlled vocabularies for the description of instruments which is an important building block in describing and identifying instrument types and their function and specifications.

Research cruises

PIDs for research cruises would be a valuable asset to the marine science community, providing the means to link cruise information (cruise track, start/end date, principle investigator etc.) to samples, funding, data and other research products generated from the cruise. Currently, we are not aware of any efforts to develop a new identifier for research cruises or any efforts to adapt current metadata schemes to encompass critical information fields related to research cruises (e.g. Vessel name, Port of entry/exit). The Rolling Deck to Repository (R2R)²⁷, a state-of-the-art US research cruise data repository, started routinely publishing DOIs for each completed cruise. Thereby, they have been able to link Cruise DOIs to related persistent identifiers where available, including ORCID iDs for members of the science party, the IGSN for physical specimens collected during the cruise, the Open Funder Registry (FundRef) codes that supported the experiment, and additional DOIs for datasets, journal articles, and other products resulting from the cruise. This approach works well under the premise that the DOI resolves to a R2R landing page which provides the necessary cruise information. In PANGAEA, partial cruise information (only related to the dataset itself) is available through the “event” label. However, cruise DOIs are currently not in wider practice by the research community, so an integration has not been attempted. Going forward, PANGAEA will implement PIDs for research cruises if these become available with metadata reflecting the complex nature of the scientific efforts or other solutions have been developed (e.g. through the O2A system at the Alfred Wegener Institute).

²⁴ <https://www.rd-alliance.org/groups/persistent-identification-instruments-wg>

²⁵ <https://github.com/rdawg-pidinst/schema>

²⁶ <https://www.pidconsortium.net>

²⁷ <https://www.rvdata.us/>

Facilities

Large-scale research facilities, depending on the information context, can be considered organizations, funders, or instruments. After discussions at STFC with facilities impact managers it was decided to use the respective existing and emerging PIDs frameworks for organizations, funders or instruments rather than agree on and promote a specific new PID type for facilities. In the STFC Open Science Portal prototype reported in D4.7, facilities are modelled as organizations with ROR or GRID PIDs assigned to them where they exist. A smaller experiment is ongoing with one STFC facility assigned with CrossRef Funder ID and some evidence that its use is growing over recent years.²⁸ There is currently no apparent indication that other facilities will follow suit. It is likely that in most research information contexts, organization identifiers will be assigned, and in more specialized contexts of instrumentation development where it is important to express parent-child relationships between facilities and instruments installed on them, Instrument PIDs for facilities may find their use.

²⁸ STFC Central Laser Facility as a funder and publications attributed to it:
<https://search.crossref.org/funding?q=100013266>

3 Disciplinary integrations

This section provides reports from six FREYA partners who have built further implementations to include new and emerging PIDs in disciplinary systems, databases and workflows. A focus across disciplines has been to incorporate ROR IDs as one type of organization ID into repository records. Some partners have used the opportunity to implement other organization identifiers at the same time, namely GRIDs and Wikidata identifiers. Another focus has been implementing funding identifiers, in the form of grant identifiers (namely those grant IDs that are assigned by individual funders) and funder identifiers (in the form of identifiers from Crossref's open funder registry (FundRef)).

To help the reader, FREYA partners have reported their implementations using the following structure where possible:

- PID resource selected
- Community served by the implementation
- The integration(s): specifications considered and achievements
- Lessons learned that partners would want to share with those wishing to undertake similar implementations

3.1 EMBL-EBI

This section outlines contributions specifically from Europe PMC at EMBL -EBI.

PID Graph resource selected

ROR IDs - organization identifiers: Europe PMC has built upon its earlier pilot integration of ROR IDs (reported in D4.4) to generate an algorithm that is able to predict ROR IDs for affiliations in existing publication records.

Community served

EMBL-EBI, which represents a life sciences community, had the following use case for organization IDs: “As an organization I want to be able to find papers published by my collaborators quickly and be able to easily find which organizations with which we collaborate.” Addressing this use case would greatly assist reporting for the EMBL annual report and funder reports. In particular, organization IDs would assist in building lists of publications authored by staff, especially those that are not found by traditional approaches e.g. via an institute’s current research information system (CRIS system) such as Converis or by using ORCID IDs for EMBL-EBI staff.

The aim was to generate an algorithm able to detect publications within EuropePMC with EMBL-EBI authors and thereby a prototype that could be extended to match all affiliations in Europe PMC records to ROR IDs.

Specifications for integration

Europe PMC arrived at the specification through the following steps:

- Determining the information required to map affiliations to ROR IDs: Affiliations appear as text strings in publications - these need to be mapped to ROR IDs. The aim was to find all papers published with an EMBL-EBI author from 2019 and 2020.
- Making use of the ROR API: We determined whether the API could return sufficiently accurate results for this mapping; It did not, so we determined the limitations of information in the ROR registry.
- Developing alternative approaches to enable mapping: this required development of a database of machine learning models capable of processing all data in the ROR registry and an algorithm able to use those models to predict potential ROR ID matches for an affiliation. As a test dataset, we made use of the previous pilot integration of ROR IDs for EMBL-EBI 2016-2018 papers (reported in D4.4).

Outcomes:

- Europe PMC’s literature database and its APIs make it easy to access affiliations in scientific papers, especially since each publication record is able to host multiple affiliations per author. The affiliation information comprises text strings that helpfully often contain geolocation of the affiliation and acronyms of the organization as well as the name by which the organization is known. The text strings can also contain information about the researcher (such as email address) or the specific unit of the organization to which the researcher belongs.
- Using the ROR API was an obvious first resort to link affiliations in text strings to ROR IDs. Its performance was tested on a dataset of EMBL papers, with particular attention given to the 6 different EMBL sites (Barcelona, Spain; Grenoble, France; Hamburg, Germany; Heidelberg, Germany, Hinxton, UK; Rome, Italy). The accuracy of the ROR API when applied to identify manuscripts with EMBL authors, was 70% which was not sufficiently reliable for our purposes. This was largely due to the fact that the city location of an affiliation is not provided in the ROR registry and thus cannot be used by the ROR API.

- A machine learning approach was developed that could incorporate the missing city geolocation to help identify EMBL papers. A significant improvement of 99.9% reliability was realised for identifying papers containing an EMBL organization. Ultimately machine learning, based on country and city information within affiliations, was combined with using the ROR API (for cases with no mapping), followed by a less selective machine learning approach. This combination gave the best results and is described in the next section.

Integration achieved

As a first step, affiliation data from ROR was collected and saved in Europe PMC's core database. This data was supplemented with information about cities that was pulled in from the GRID database²⁹; information about cities is not currently supported by ROR but there is fortunately a 1:1 match between ROR and GRID records. Next an algorithm was built to use this database to train models for each organization and use those models to predict ROR IDs in affiliation text strings. The algorithm to match ROR IDs to affiliation strings in Europe PMC records was then built using location information in the affiliation: country, city or regions were detected using dictionaries built by the Geonames Database³⁰ that support natural languages (for example, Turin is Torino in Italian). A training set of papers was used: these are papers officially identified as EMBL-EBI publications for 2015-2018 for which ROR IDs had been manually identified.

Initially, the aim was to obtain matches to a small number of ROR IDs and then to improve precision. ROR IDs with the best result were returned ordered with their corresponding scores. If the machine learning approach failed to find a corresponding ROR ID then a call to the ROR API was made and machine predictions tried again. Ultimately, a heuristic approach was used to weight scores between the ROR API and machine learning approach. Figure 1 shows the results obtained from the mapping algorithm towards the end of the training procedure. Figure 2 provides a diagrammatic summary of the approach taken by Europe PMC to map affiliations in publications to ROR IDs.

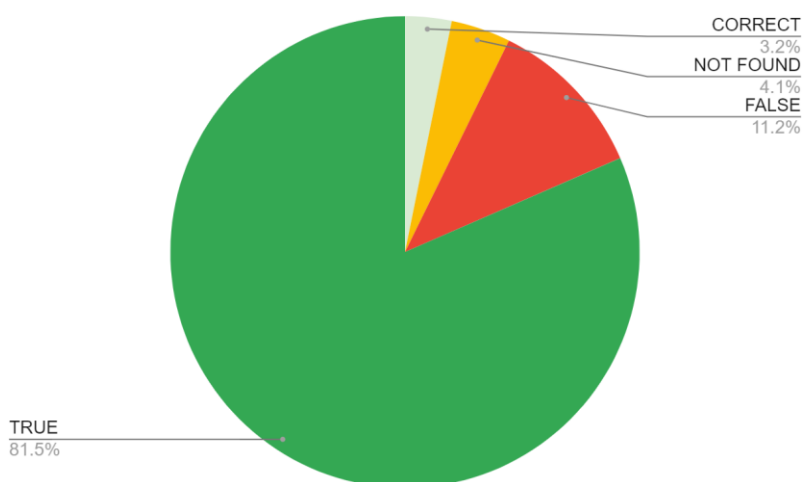


Figure 1 Algorithm performances, last version on 10/06/20: Optimisation of the mapping algorithm (using the 2015-2018 dataset) is almost complete. Performance (current as of 06/2020) is good on a set of 1656 affiliations: 81.5% were predicted as correct ROR IDs (TRUE), 11.2% were wrongly predicted (FALSE), 4.1% were not assigned to a ROR ID (NOT FOUND) and in the end 3.2% seems to have the correct ROR ID but the annotator could not be sure about it (CORRECT).

²⁹ The GRID database: <https://www.grid.ac/>

³⁰ The Geonames database: <http://www.geonames.org/>

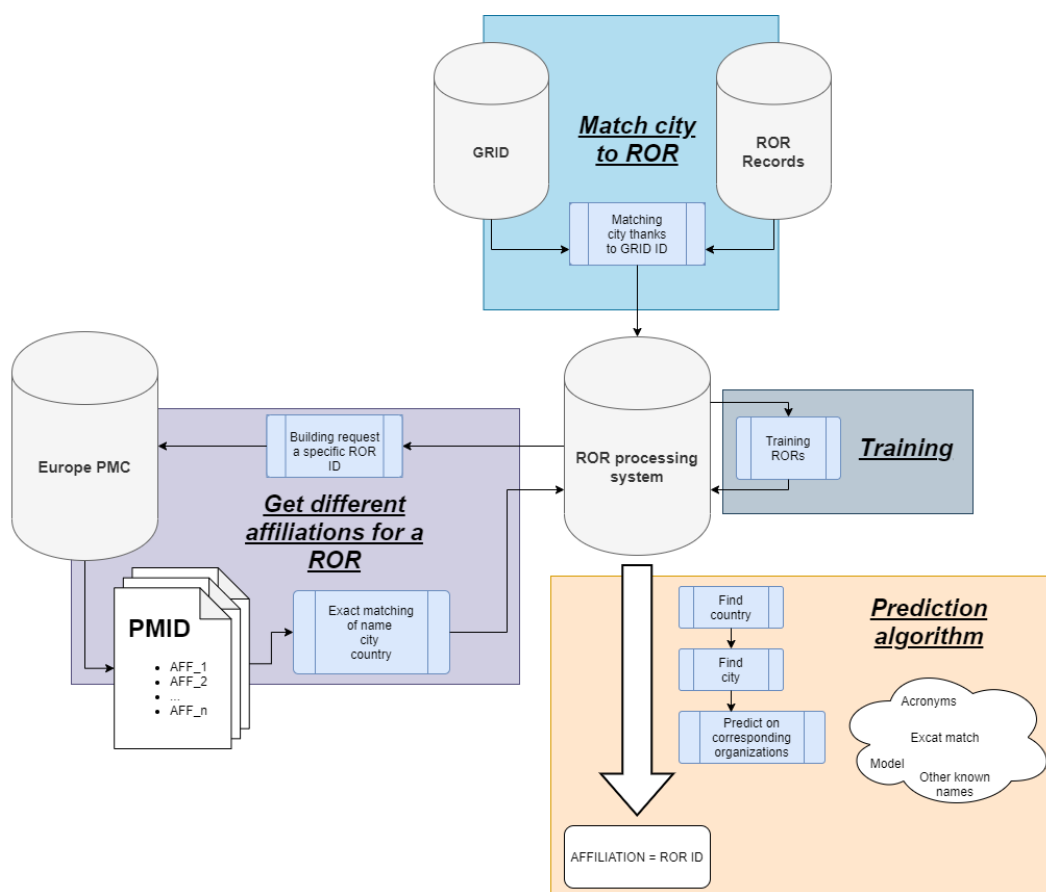


Figure 2 Database creation and plan for algorithm: This diagram represents how the database (ROR processing system) was set up for the prototype. The first step (blue, top) involves merging the city data from GRID to the ROR IDs. A city field has been added to each ROR ID. In step 2 (purple), model training data were needed - here using EuropePMC records (indicated by PubMed Identifiers or PMIDs), an exact match of the name of the organization, the city and the country datasets has been created for each ROR ID. Those datasets contain from 0 to more than 8000 different strings for each ROR ID. Once there are enough strings (at least 100) for a ROR ID, the prototype trains a simple model (grey square) based on this dataset (comprising correct matches) and strings from other organizations (wrong matches). In the end the newly created database (ROR processing system) contains a dataset of different strings for each organization retrieved for Europe PMC records - in addition, where there are sufficient different affiliations, the model should be able to predict whether a new string corresponds correctly to a given ROR ID. The resulting algorithm (orange) will find the location information and through processes involving exact match, acronyms, trained model and other known names, will predict whether the string corresponds to a ROR ID.

Resulting integration: All papers published with an EMBL-EBI author from March 2019 to March 2020 have been tagged with the corresponding ROR ID (<https://ror.org/02catss52>) and this information is currently live within Europe PMC. See Figure 3. In addition the official EBI publication lists from 2015 to 2018 have been tagged with predicted ROR IDs (all affiliations for all authors).

The screenshot shows the Europe PMC search interface. At the top, there is a navigation bar with 'About', 'Tools', 'Developers', and 'Help'. Below this is a search bar containing the query 'org_id:ror.org/02catss52'. A green banner below the search bar says 'Search worldwide, life-sciences literature' and includes a link for 'Coronavirus articles and preprints' and search examples like '"breast cancer" Smith'. There are tabs for 'Recent history' and 'Saved searches'. The main content area is divided into two columns. The left column contains filters: 'Search only' (1-25 of 1,033 results), 'Type' (Research articles (930), Reviews (103), Preprints (0)), 'Free full text' (Free to read (862), Free to read & use (733)), and 'Date' (2020 (48), 2019 (258), 2018 (250), Custom date range). The right column displays search results. The first result is 'A comprehensive and comparative phenotypic analysis of known phenotypes' by Kollmus H, Fuchs H, Lengger C, Haselimashadi JA, Amarie OV, Becker L, Beckers J, Calzada-Wack J, Kuckuk P, [...] Hrabě de Angelis M, published in Mamm Genome, 31(1-2):30-48, 14 Feb 2020. The second result is 'Comparing Cryo-EM Reconstructions and Validations of Protein Structures' by Joseph AP, Lagerstedt I, Jakobi A, Burnley T, Patvardhan S, published in J Chem Inf Model, 60(5):2552-2560, 11 Feb 2020. Both results include citation counts and 'Add to export list' links.

Figure 3 Europe PMC's resulting integration: A screenshot from Europe PMC's website showing a list of publications with EMBL-EBI authors, returned using the ROR-ID for EMBL-EBI in the search query: `ORG_ID:ror.org/02catss52`.

EMBL-EBI work for STFC:

On learning of Europe PMC's algorithm and database for matching affiliation strings to ROR IDs, FREYA partner STFC wished to replicate this approach for their own data.

STFC provided a list of DOIs corresponding to 40,000 publications in STFC repositories. Europe PMC agreed to apply its machine learning algorithm to map these DOIs to PubMed Identifiers (PMIDs) to ROR IDs, producing a Google data studio report - see a screenshot in Figure 4.

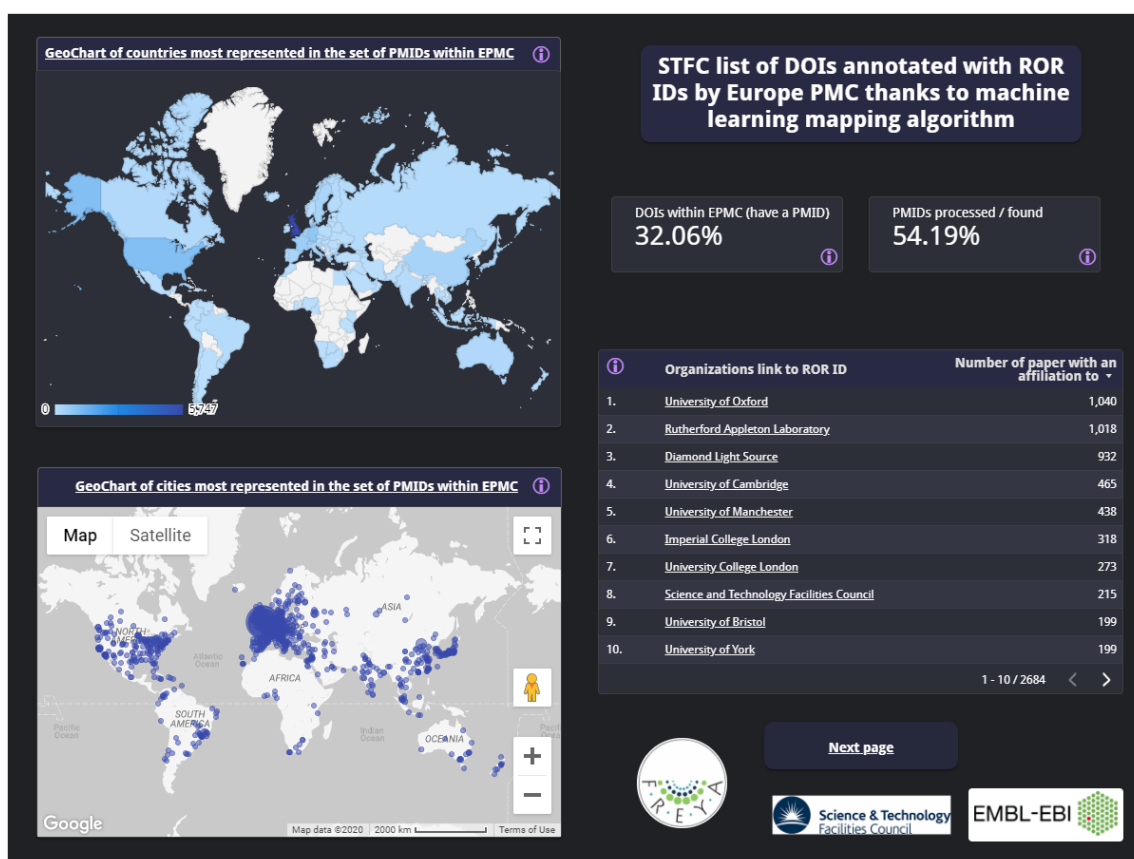


Figure 4 A screenshot of results obtained after applying the machine learning prototype to STFC's DOIs: 32% (~12 8000) of the DOIs provided by STFC have a corresponding PMID in Europe PMC. The "heatmap" (top left) presents countries most represented in the STFC dataset while the second map (bottom left) displays the cities most represented. Note that the prototype requires a processing time of about 5 seconds per affiliation and this is reflected in the report by the statistic displaying "PMIDs processed vs PMIDs found" (here 54.19% of the PMIDs found in the 40,000 DOIs from STFC have been processed by the prototype). The table (bottom right) displays most represented organizations among STFC Publications ranked in descending order. Each organization name is hyperlinked to its corresponding entry in the ROR registry (e.g. University of Oxford: <https://ror.org/052gg0110P>).

Lessons learned

This is the information imparted to STFC ahead of the collaborative work that was undertaken.

- For mapping affiliation strings to ROR IDs, it is essential to use cities as it distinguishes organizations that have different sites in the same country (e.g. currently the ROR API is not able to differentiate between EMBL Heidelberg and EMBL Hamburg). The downside of using cities is that for some affiliations, establishing the correct location can take a lot of time - there are many countries/cities regions to check. A possible solution for this might be to use the geocoding API from Google which will return geolocation information for an affiliation such as country, city, regions, etc.
- Once the geolocation of an affiliation is known then we suggest the best approach is to use a machine learning algorithm supplemented if necessary by the ROR API. The machine learning approach can predict the correct ROR ID for an affiliation even if the affiliation is misspelled in a text string. Indeed the machine learning prototype built by EMBL-EBI to assign ROR IDs to affiliations provides good results (81.5% accuracy).
- Small affiliations are not processed by the machine learning approach e.g. "Gastroenterology unit", since there is not enough information to assign a ROR ID with a good reliability. Time-consuming and inaccurate predictions can also arise when location information is missing.

- Europe PMC's machine learning prototype utilizes text strings from publications that are assigned both DOIs and PMIDs³¹. Europe PMC is currently exploring possible ways to make the data public going forward.

³¹ From Deliverable 3.1: "Publications can be referred to by several equivalent PID types. Taking the example of PMID/PMC/DOI for *publications*: a journal article (with a DOI that is allocated by a publisher and metadata registered with the DOI registry, Crossref) may also have its abstract indexed by PubMed if it is biomedical, and will be assigned a PMID (PubMed reference number) by the National Library of Medicine (USA). The PMID refers to the metadata record in PubMed. If the full text version of the same journal article (already identifiable via a DOI and PMID) is indexed in EuropePMC/PMC (an archive of full-text journal articles) then it will be assigned a PMC identifier, which refers to the full text version in EuropePMC/PMC. The PMID/PMC/DOI identifiers are equivalent in that they refer conceptually to the same article, but the specific instances are different."

3.2 DANS

PID Graph resource selected

Specific Grant IDs and (Funding) Project IDs, but connected to other entities in the PID-Graph.

Community served

The Dutch multidisciplinary NARCIS community in general, and its funding agencies more specifically. Results are presented to different stakeholders and in a presentation about the NARCIS PID-Graph.

The portal NARCIS (www.narcis.nl) is an aggregation of Dutch Research Information. NARCIS contains the metadata from all the Institutional Repositories in the Netherlands with publication metadata, metadata of datasets, metadata of research projects, researchers and organizations. Dutch universities and other research organizations are all participating in NARCIS. NARCIS also contains the project grants (project - or programme descriptions) of the Dutch Research Council and European Commission (EC) where Dutch organizations are involved in.

NARCIS aggregates information from more than 48 different institutional repositories and 23 archiving systems. In addition to the aggregated metadata from these repositories, NARCIS contains information from Current Research Information Systems (CRISs), information about Dutch research organizations, researchers and experts, and research projects.

Specifications for integration

Within D4.6 DANS worked on the question of whether it would be possible to connect grant descriptions to different types of publications and datasets on the basis of Persistent Identifiers (PIDs). A persistent identifier (PID) is a long-lasting reference to a document, file, web page, or other object, which is resolvable to the current location on the web.

Because NARCIS contains grant information, as well as publications and datasets, it provides a good testing ground to look into this matter and determine potential challenges. The PIDs necessary to connect this information, including person IDs and organization IDs are all part of the NARCIS-PID Graph.

To implement grant-IDs and link the funded projects to publications and datasets that resulted from those projects, the namespace "info:eu-repo/grantAgreement/" is used and this ID is included in the NARCIS- PID Graph.

One of the use cases high on the priority list of NARCIS, is the relation between grants and the results of the funded projects: publications and datasets. Different stakeholders of NARCIS want to get insights in the results of a certain grant, in terms of the different kinds of publication types (article, book, report etc) and datasets. Within this task, DANS implemented this use case for European Commission (EC) grants and set up a pilot for a grant given by the Dutch Research Council (NWO).

Current Status: NARCIS is ready to support Grant IDs and connect these to (funding) projects. The situation described in "integration achieved" is live in our production environment. Not all sources included by NARCIS can offer Grant IDs yet, however it is expected that the illustrations built in the production environment of NARCIS will fuel implementation in the different source systems.

Integration achieved

European Commission (EC) funding

One of the purposes of NARCIS is to propagate metadata of Dutch research institutes to other information services around the globe. Among those information services is OpenAIRE. Dutch research institutes can submit their metadata in a number of different formats to NARCIS, and NARCIS propagates this information

in a standardized way to OpenAIRE. This metadata also includes publications and datasets that are a result of an EC-funding and NARCIS wants to include these relations in the metadata and thereby also in the PID Graph. In addition, these relations are presented in the web-interface (narcis.nl) and propagated to other services like OpenAIRE.

To stimulate the use of the EC grant IDs in the Dutch scientific community, we built a few showcases in NARCIS which were used in presentations and communication to the different communities. These activities led to a growth of 850 publications with a EC grant ID in the metadata that could then be linked to the right grant and project description.

A great example is the EC project “Realising an Applied Gaming Eco-System - RAGE”³²; see Figure 5. Using the namespace for EC-funding, this project is linked to 91 different publications by including the right grant ID in the metadata.

The publications can be filtered by publication type, year, accessibility and source. In PID Graph terms it is possible to present the total number of publications from a certain grant, or present the total numbers of open, closed, and restricted access publications. It is all linked by the grant ID “info:eu-repo/grantAgreement/EC/H2020/644187”.

RESEARCH
REALISING AN APPLIED GAMING ECO-SYSTEM - RAGE

<p>Main</p> <p>Publications (91)</p> <p>Update content</p> <p>Type</p> <ul style="list-style-type: none"> Other (50) > Conference paper (23) > Article (14) > Report (2) > Book (1) > All types + <p>Date</p> <ul style="list-style-type: none"> 2020 (1) > 2019 (7) > 2018 (12) > 2017 (26) > 2016 (34) > All dates + <p>Accessibility</p> <ul style="list-style-type: none"> Open Access (46) > Closed Access (41) > Restricted Access (4) > <p>Source</p> <ul style="list-style-type: none"> Open Universiteit Nederland (91) > 		<p>Title Realising an Applied Gaming Eco-system - RAGE</p> <p>Period 02 / 2015 - 02 / 2019</p> <p>Status Current</p> <p>Research number OND1358796</p> <p>Data Supplier Horizon 2020</p> <p>Funded by info:eu-repo/grantAgreement/EC/H2020/644187</p> <p>ABSTRACT</p> <p>The EU based industry for non-leisure games (applied games) is an emerging business. As such, it's still fragmented and needs critical mass to compete globally. Nevertheless its growth potential is widely recognised and even suggested to exceed the growth potential of the leisure games market. RAGE will help to seize these opportunities by making available 1) an interoperable set of advanced technology assets tuned to applied gaming 2) proven practices of using asset-based applied games in various real-world contexts, 3) centralised access to a wide range of applied gaming software modules, services and resources, 4) an online social space (the RAGE Ecosystem) that arranges and facilitates collaboration that underlie progress and innovation, 5) workshops and online training opportunities for both developers and educators, 6) assets-based business cases that support the games industry at seizing new business opportunities, and 7) a business model and launch plan for exploiting the RAGE Ecosystem beyond the project's duration. Intermediary organisations and education providers anticipate a wider exploitation of RAGE results among their end-users, which add up to over 1 million, and through disseminating RAGE in their partner networks. The game companies in RAGE anticipate adding RAGE-based products to their portfolio, in order to improve their competitive advantage by opening a new product line for applied games and developing new revenue streams. Actual deployment of RAGE results will generate direct impact on the competitive positioning of the few thousand of European SMEs in the Applied Games market. Impacts from RAGE will be visible in terms of fulfilling new client needs by quicker and more challenging methods of skills acquisition, enabling new business models based on the usage of the assets repository and the Ecosystem, and in the strengthening collaboration across the entire Applied Games value chain.</p> <p>RELATED ORGANISATIONS</p> <p>Secretariat > Faculty of Educational sciences (OU)</p> <p>Financier > European Commission</p> <p>RELATED PEOPLE</p> <p>Project leader > Prof.dr. W. (Wim) Westera</p> <p>RELATED RESEARCH (UPPER LEVEL)</p> <p>Funding Programme > Horizon 2020</p>
---	--	--

Figure 5 NARCIS page of the EC-grant “Realising an Applied Gaming Eco-system - RAGE”

Dutch Research Council (NWO) funding

It is a long-standing wish to relate publication and dataset to grants from the National Research Council (NWO). Currently this work is done manually by asking researchers to add their publications to the

³² <https://www.narcis.nl/research/RecordID/OND1358796>

application that supports the administrative process around the NWO research grant. Together with the Donders Institute for Brain, Cognition and Behaviour and the Radboud University we set up a pilot to explore this use case for the research grant called “Language in Interaction”³³; see Figure 6. This programme is divided into five different so called “Big Questions” with one or more research projects associated with them. In this pilot Donders Institute, Radboud University and DANS aimed to sort out a number of questions. Most importantly, we wanted to know whether it is possible to connect the grant, the different projects, the publications and the datasets on the basis of PIDs and whether the current PIDs or IDs for grants and projects are sufficient. We wanted to assess how such a workflow can be organized and what problems we would encounter along the way. In particular we wanted to assess whether the namespace “info:eu-repo/grantAgreement” is usable for this process. This namespace is used by OpenAIRE for EC grants and in this pilot we wanted to explore if this can also be used for grants from the Dutch Research Council (NWO), or whether it would be better to adopt to grant IDs from CrossRef.

The ultimate goal of this pilot was to see whether publications and datasets could go through the entire workflow and systems from the participating organizations and then have their metadata added to NARCIS and linked to the NWO grant “Language and Interaction”.

For this pilot most of the needed PIDs were available. Persons can be identified with ORCID iDs or ISNIs, and organizations with RORs. Publications and datasets are identifiable through DOIs or Handles. However, for grants and projects the situation was not that obvious. For projects, the pilot used the NARCIS project IDs, which are persistent and sustained for more than fifty years, but these identifiers do not meet all the requirements of a persistent identifier, mainly because they are not actionable on the web and do not resolve to a landing page. There is thus still a need for an Open Persistent Identifier for projects!

Based on the OpenAIRE grant ID, the pilot used the current NWO grant ID. This grant ID is not resolvable, and there is also no guarantee that it will be sustained in the future. In that sense it is questionable if this ID meets the criteria for a persistent identifier. However, the ID, together with the name space, is globally unique and can be traced to the source, namely NWO. The pilot did use the “info:eu-repo/grantAgreement” namespace, according to the OpenAIRE guidelines³⁴, with the extension “/NWO/Gravitation/024.001.006”. According to the specifications, these last three fields are reserved for /Funder/FundingProgram/ProjectID/

The pilot clearly shows that in the Netherlands there is still a need for the implementation of “real” grant PIDs and PIDs for projects, programmes or other types of research descriptions and CrossRef might be a good alternative for grant IDs.

Technically it is rather simple to aggregate the publications and datasets and connect them to a project on the basis of a grant ID in the PID-Graph. However, there are challenges in organizing this work. Lack of clarity on the use of grant IDs, the different formats used in different systems and practical problems in assigning PIDs to the metadata, makes it more complicated than just the technical implementation.

Although the pilot was limited, the different systems, using different formats provided us with insight into the problems that can arise. NARCIS received datasets directly from The Language Archive owned by Donders Institute and datasets from the Radboud University. Both are using different formats and not all the formats do support grant information. Publications related to “Language in Interaction” are stored in institutional repositories owned by various universities. Not all systems were able to assign the grant ID “info:eu-repo/grantAgreement/NWO/Gravitation/024.001.006” to those publications.

It has also been found that from the viewpoint of researchers, assistance is needed in assigning a grant ID into the metadata. These grant IDs are often unknown, and applications should support this process by

³³ <https://www.narcis.nl/research/RecordID/OND1366397>

³⁴ https://guidelines.openaire.eu/en/latest/literature/field_projectid.html

assigning a grant ID automatically when a researcher selects the program “Language in Interaction”, which is known.

RESEARCH
LANGUAGE IN INTERACTION

Main

Publications (7)

Datasets (3)

Update content >

Title	Language in Interaction
Period	2013 - 12 / 2023
Status	Current
URL	> https://www.languageininteraction.nl
Research number	OND1366397
Funded by	info:eu-repo/grantAgreement/NWO/Gravitation/024.001.006

ABSTRACT

Human language is the most powerful communication system that evolution has produced. It is the basis of culture and social life. It comes in many forms (> 6000 languages today). At the same time it is deeply rooted in the neurobiology of the human brain. The overarching quest of the programme is to account for, and understand, the balance between universality and variability at all relevant levels of the language system and the interplay with different cognitive systems, such as memory, action, and cognitive control. To achieve this, Language in Interaction brings together researchers from eight universities and one research institute within the Netherlands to understand this unique capacity in its full glory.

RELATED ORGANISATIONS

Secretariat	> Donders Institute for Brain, Cognition and Behaviour (RU)
Financier	> Netherlands Organisation for Scientific Research - NWO (NWO)

RELATED PEOPLE

Project leader	> Prof.dr. P. (Peter) Hagoort
Contact person	> Dr. W. (Wendy) van Ginkel

RELATED RESEARCH (UPPER LEVEL)

Funding Programme > Gravitation

RELATED RESEARCH (LOWER LEVEL)

- > Big Question 1: The nature of the mental lexicon: How to bridge neurobiology and psycholinguistic theory by computational modeling?
- > Big Question 2: What are the characteristics and consequences of internal brain organization for language?
- > Big Question 3: Creating a shared cognitive space: How is language grounded in and shaped by communicative settings of interacting people?
- > Big Question 4: Variability in language processing and language learning: Why does the ability to learn language change with age? How can we characterize and map individual language skills in relation to population distribution?
- > Big Question 5: The inferential cognitive geometry of language and action planning: Common computations?

Figure 6 NARCIS page of the NWO-Grant “Language in Interaction”

In conclusion, the namespace “info:eu-repo/grantAgreement” can be used as grant ID, but it would be far better if an open and global grant identifier were available, preferable resolving to a landing page with a description of the grant hosted by the funder. This infrastructure of course would also need to be persistent!

Lessons learned

- Technically the implementation is actually rather simple.
- However, on the organizational level there is a huge challenge. In this pilot a dataset, or publication, goes through a rather complex workflow of systems and it takes some effort before these systems can process the necessary information. Systems need to assist a researcher to get a grant number in the metadata and different systems need to bridge different formats
- There is a need for more uniformity among grant PIDs and PIDs for research projects or programmes. Universities need to support the OpenAIRE grant ID, or for example CrossRef Funding IDs as these PIDs play a crucial role to let different systems throughout Europe interoperate.
- Once requirements are met, the potential benefits are huge. Funders can retrieve metadata of publications and datasets quite easily on the basis of the right grant PIDs. In addition, PIDs of participating organizations and researchers can be retrieved as well. This actually is a practical use-case for the PID Graph in action.
- In the Netherlands there is still a debate about which grant ID will be used. Within this use-case EC-grant ID was used because this is mandatory for publications as an outcome of an EC-funded

project. It is expected that Crossref grant IDs and EC-grant IDs will be used both. Irrespective of the grant ID, however, the principles of the described use remain the same. The workflow is in production and a change of grant ID will not affect the implemented solution very much.

3.3 British Library

Community served

The British Library's Shared Research Repository is a multi-tenanted repository for cultural heritage organizations in the UK. At present, six museums and other heritage organizations use the repository service, administered by the British Library, which provides individual research repositories and a shared searchable layer. These repositories are designed to provide a unified location for the research conducted within these organizations to be found. The repository service, based on the Samvera Hyku platform³⁵, was launched as a pilot in October 2019 and will go into full service in January 2021. The content of these shared repositories includes outputs from the research conducted by the staff of these cultural heritage organizations, the students who undertake collaborative doctoral studies with these organizations and university and doctoral student placements.

The UK's index of doctoral theses, EThOS, is administered by the British Library, and is due to be migrated to a new platform in the coming years, which will likely be the same platform as the research repository. Therefore these developments will integrate with that replatforming when it takes place in 2021 onwards. EThOS has a broader audience than the Shared Research Repository as it contains theses from all fields of research and is used by researchers but also by those whose thesis is indexed there.

The developments described below will be available to other users of the Samvera Hyku platform via the outputs of the Advancing Hyku project³⁶. Ubiquity Press, the British Library's development partner for the repository, will make this code available in a form compatible with out of the box Hyku installations, meaning this work can be used by the whole of that community.

PID Graph Resource

In this piece of work, we aim to integrate emerging PID types into the British Library's Shared Research Repository. The PID types selected are PIDs for organizations and funding. Many stakeholders using the British Library's repository are interested in gaining a better picture of organizational collaboration across cultural heritage organizations and universities. Many UK cultural heritage organizations are what are known as independent research organizations³⁷ and are eligible to receive funding from the UK's national funders as well as other philanthropic organizations, integrating funder identifiers will improve the tracking and visibility of the outcomes of that research. While these developments will only be available for direct use by UK cultural heritage organizations who can use the repository service, they provide a general use case for any organization wishing to increase the number of identifiers included in a repository implementation. With regard to the eventual migration of EThOS, user story #45, User Stories for Funding PIDs, still applies, relating to tracking PhD outcomes³⁸.

Specification of changes and their importance

In integrating support for organizational and funder identifiers, it was deemed necessary that any integration include the capacity for this metadata to be included in any metadata submitted to DataCite on the minting of a DOI, so that it is included easily in the PID Graph. In making determinations about which IDs should be included, it was determined that including a broad base of organizational IDs was important to future proof the repository. Additionally usability for both repository staff and users was considered,

³⁵ <https://hyku.samvera.org/>

³⁶ <https://advancinghyku.io/>

³⁷ <https://www.ukri.org/funding/how-to-apply/eligibility/>

³⁸ User Stories for Funding PIDs: <https://www.pidforum.org/t/user-stories-for-funding-pids/102>

which is why not all identifiers will be immediately visible on the front end. While this enhanced metadata will be included in all new records to the repository, it will only be applied to existing records if they are edited, with the exception of the funder field, described below.

Integration achieved

The British Library worked with their development partner, Ubiquity Press, to make the developments to the repository service. Namely, Organisational Creator and Contributor fields were altered. Previously these properties had two fields, the Organisational Creator/Contributor name and the Organisational Creator/Contributor ISNI. The changes mean that additional fields are now available to support ROR, GRID and Wikidata IDs, as illustrated in Figure 7. ROR IDs are now visible on the front end, as ISNIs were previously, as shown in Figure 8. GRID and Wikidata IDs will not be visible to end users of the repository but all identifiers will be included in the metadata submitted to DataCite when a DOI is created. As the repository has a mediated deposit process, the standard of metadata added to records is high and identifiers are included wherever they can be reliably retrieved. The fields will also include pattern based validation to ensure that the fields are presented as resolvable URLs.

Creator name type required

The person or group responsible for the work. Usually this is the author of the content.

Creator organisation name

Please complete this field.

Creator organisation ROR

Creator organisation GRID

Creator organisation Wikidata

Creator ISNI

[✖ Remove](#) | [Add another](#)

Figure 7 A screenshot of the new upload form in British Library's Shared Research Repository. Users can now enter a variety of identifiers including ROR, GRID, Wikidata, in addition to the existing ISNI. ROR and ISNI identifiers are displayed on the front end.

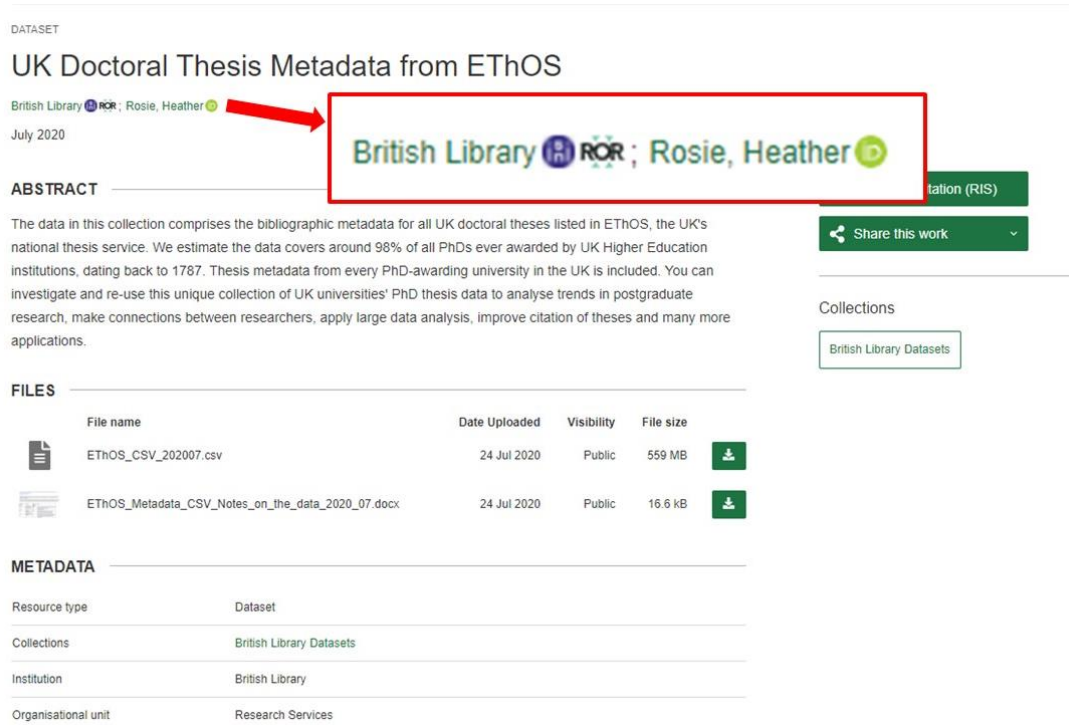


Figure 8 A screenshot showing a ROR ID for a Creator in the front end of the Shared Research Repository

For the Funder field in the upload form for the repository, which was a dropdown list of 40 funders, provided by the cultural heritage organizations and manually populated by the development partner, this will now become an auto-suggesting field which will suggest funder names when a user starts to type in the name of a funder, see Figure 9. This will then populate with the Crossref funder ID, the ISNI and ROR identifier. Where no funder name is suggested this can be typed in manually and the ROR and ISNI identifiers populated. It is also now possible to include multiple funder project references, where there was previously only one field, and these references are associated with an individual funder.

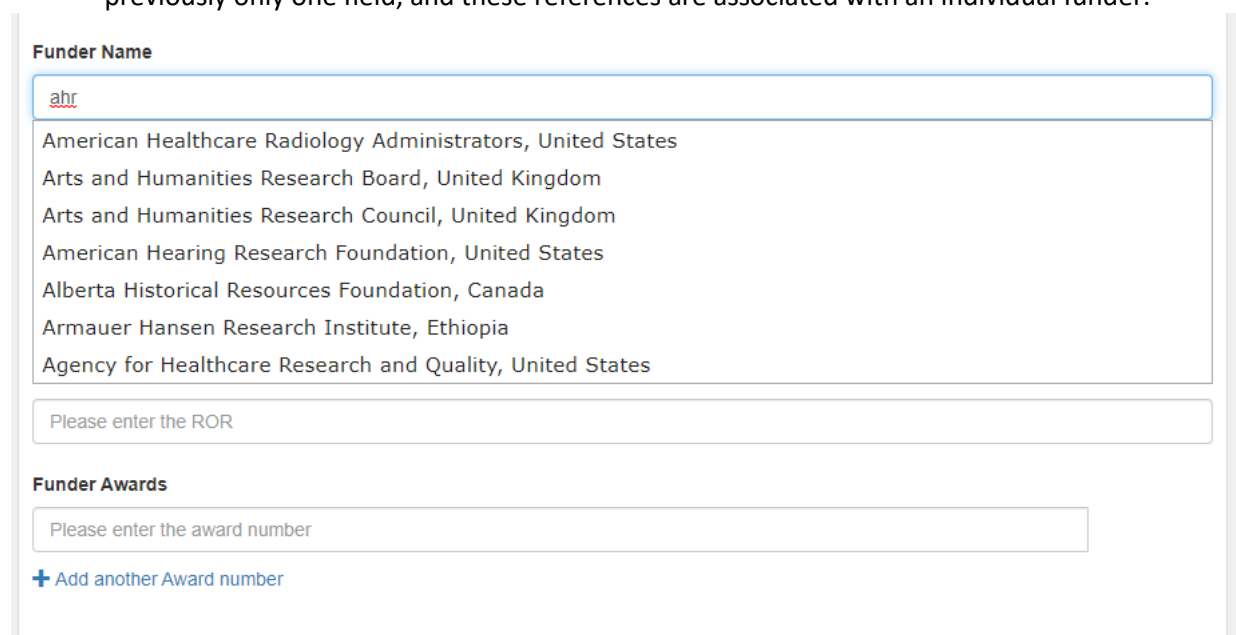


Figure 9 An image of the autopopulating list based on the Crossref funder registry in upload form of the British Library’s Shared Repository

All metadata and identifiers will be included in the submission to DataCite when DOIs are created. ROR and ISNI icons for the funder are visible on the front end with the new configuration of funder and funder project reference, see Figure 10.

Collective Re-Excavation and Lost Media from the Last Century of British Prehistoric Studies

Wexler, Jennifer ; Bevan, Andrew ; Bonacchi, Chiara ; Keinan-Schoonbaert, Ad ; Pett, Daniel ; Wilkin, Neil 

24 April 2015

ABSTRACT

There are thousands of forgotten archaeological archives hidden away in repositories all over the world, lost worlds where many scholars have toiled away for years, trying to record every detail and bit of information available about rare and precious archaeological objects in an attempt to bring order and understanding to an almost incomprehensible past. This paper discusses how these archives can be approached through Huhtamo's definition of media archaeology as a 'historically-attuned enterprise' that involves 'excavating forgotten media-cultural phenomena', focusing on the MicroPasts digitization project. It is shown that greater utilization of digital media simply changes and extends the terms of engagement, accessibility, and flow of information from antiquated archaeological archives to the community and back again.

 Download citation (RIS)

 Share this work

FILES

File name	Date Uploaded	Visibility	File size
 Adl_aam_collective.docx	14 Sep 2020	Public	2.98 MB 

METADATA



Resource	Funder	Arts and Humanities Research Council  (Award number: AH/M00953X/1)
Institution		
Organisational unit	Digital Scholarship	
Funder	Arts and Humanities Research Council  (Award number: AH/M00953X/1)	
Journal title	Journal of Contemporary Archaeology	

Figure 10 A screenshot highlighting the funder field of a record in the British Library's Shared Research Repository

A "rake" task was performed to make all existing records in the repository compatible with this new data model. All records which have funding information will be manually updated to make sure this new format funding information is displayed on the front end.

To support the potential migration of EThOS metadata, the "current HE institution" field in the thesis template was updated to include ISNI and ROR identifiers. This field is mandatory for all thesis records and is based on a controlled list. As the list is controlled the IDs will be included as a table within the repository.

Lessons learned

This work was implemented successfully but there were a few things which were noticed during the implementation which may be of benefit to others attempting similar implementations. The consideration of how to format the identifiers took some time and it was not possible to get clear guidance on how to format identifiers such as GRID and Wikidata in the DataCite metadata schema. There was also a lack of clear guidance from identifier providers on how to display identifiers, e.g. there is no clear indication on the ROR website how and when the logo should be used.

We initially had some issues with the funder registry look up in that the way it self-selects based on typing is not always intuitive or expected. It was determined this was due to the way the Crossref sorts the results when called from the API based on meaningful words in the Funder name.

3.4 CERN

PID Graph resource selected

This work is pertaining to ORCID iDs and ROR IDs for the CERN Open Data portal, and PIDs for funding, grants and organizations for the Zenodo platform.

CERN Open Data portal

The CERN Open Data (COD)³⁹ portal is a “big data” open-access repository currently containing over 2PB of particle physics data and accompanying code/software and documentation produced through the research performed at CERN. While ORCID iDs already existed in COD (see Deliverable 4.1), it was determined that certain improvements needed to be made. Making the portal ROR “friendly” is an enhancement which is part of the larger effort of adding support for organization PIDs in CERN services that started with the work presented in Deliverable 4.4.

Zenodo

Zenodo⁴⁰ is a generalist open research repository hosted by CERN and commissioned by the European Commission (EC) through the OpenAIRE project to support the Open Data and Open Access movements in Europe launched in 2013. Zenodo includes support for various mature PID systems already (e.g. DOIs, ORCID iDs, ISBNs) and integrations for new PIDs are considered equally important to keep up with the wider open science community, which in this case involves PIDs for funding and grants. Furthermore, CERN, together with 18 international partner institutions, currently builds a turn-key Open Source research data management platform called InvenioRDM, for which the main purpose is to enable reproducibility and reuse of research artifacts⁴¹. Support for ROR will be part of the development of InvenioRDM and as a result will then be available on Zenodo as well.

While these developments on the Zenodo service have not been the result of FREYA work and Zenodo is not one of CERN’s pilot applications in FREYA, it is important to point them out as they pertain to work on the exact same topics that FREYA addresses and it is a way of giving an overview of PID-related work at CERN more broadly.

Community served

The presented developments are part of the overall effort at CERN to integrate new and emerging PID types and to continuously make already-established PID integrations better.

The CERN Open Data portal is an open science repository that is used by researchers, educators and students of multiple levels. While it currently holds outputs from the CERN community (i.e. results from the main LHC experiments and other experiments at CERN), the materials published on the platform can be of value for anyone working on the field of High-Energy Physics in general or even more broadly in the context of data science and machine learning.

Zenodo as a generalist research repository serves researchers from all disciplines across the globe. It is free to use and accepts all kinds of research artifacts, which makes it widely adaptable to a very broad community. InvenioRDM is a repository application that can be used by anyone to run a service similar to Zenodo. As it is currently developed by CERN and external partners such as Northwestern University and openly available, its community is CERN (mainly Zenodo) and current and future service owners across the

³⁹ <http://opendata.cern.ch/>

⁴⁰ <https://zenodo.org/>

⁴¹ <https://inveniosoftware.org/blog/2019-04-29-rdm/>

globe. Thus, every feature of InvenioRDM has the potential to be easily implemented by numerous research repositories.

Integration achieved

CERN Open Data portal

In terms of the ORCID enhancements on the portal, this work included retroactive addition of missing ORCID iDs to old records. Previously, ORCID iDs had been added to a small number of records as a first step and this is the completion of that effort. Another task was making the ORCID iDs actionable through the user interface as in the past it was only possible to see the ORCID iDs through exporting the record metadata which is not ideal for the user.

More than 200 (not unique) additional ORCID iDs were added to COD records. There was a very small number of authors (under 10 names) that could not be matched with ORCID iDs from past records which is something that will need to be resolved in a different way in the future.

As far as adding support for ROR IDs on the portal, the schema was adjusted to make it possible to add ROR IDs in the metadata (Figure 11).

```
"authors": [  
  {  
    "affiliation": "Helsinki Inst. of Phys.",  
    "rorid": "01x2x1522",  
    "name": "Lassila-Perini, Kati",  
    "orcid": "0000-0002-5502-1795"  
  }  
]
```

Figure 11 Metadata of a record with the added ROR field for authors

A handful of ROR IDs were added to a few selected records as an MVP and it is intended that this implementation will be extended in the future. As with the ORCID iDs, the ROR IDs were made visible on the detailed record. The user can navigate to the ORCID or ROR record of the author or organization by clicking on the logos next to the names and affiliations. Figure 12 shows the user interface before and after this development.



Figure 12 Example of a CERN Open Data portal record displaying ORCID and ROR IDs (bottom) and how it looked like before this work was done (top)

Zenodo

Zenodo already integrated funder DOIs connected to the Crossref Funder Registry⁴², which can be connected to internal grant IDs; non-actionable strings that vary across funders. The Crossref Funder Registry indexes these internal grant IDs and links them to the funder. These integrated grant IDs are not PIDs and must not be confused with the grant DOIs linked to the global grant identifier system introduced in Deliverable 3.1⁴³. In detail, the Zenodo metadata mask for every upload includes the field “Funding”, where OpenAIRE-supported funders⁴⁴ can be added (see Figure 13: Record upload mask in Zenodo). As Zenodo is integrated in OpenAIRE’s reporting system, the selected funding agency will be informed that the uploaded research artifacts are now (openly) available on Zenodo. The grant number of the project can be specified in a separate field. Funders that are currently not supported by OpenAIRE can be specified in the “additional notes” field, but they will not be notified by this upload.

⁴² <https://www.crossref.org/services/funder-registry/>

⁴³ <https://doi.org/10.5281/zenodo.3554255>

⁴⁴ Supported funders are: Australian Research Council (AU), Austrian Science Fund (AT), European Commission (EU), European Environment Agency (EU), Academy of Finland (FI), Fundação para a Ciência e Tecnologia (PT), Hrvatska Zaklada za Znanost (HR), Ministarstvo Prosvete nauke i Tehnološkog Razvoja (RS), Ministarstvo Znanosti Obrazovanja i Sporta (RS), National Health and Medical Research Council (AU), National Institute of Health (US), National Science Foundation (US), Nederlands Organisatie voor Wetenschappelijk Onderzoek (NL), Research Councils (UK), Science Foundation Ireland (IE), Wellcome Trust (UK)

Figure 13 Record upload mask in Zenodo

The specified grant ID and the funder ID are searchable objects in the metadata (see Figure 14). The selected funder is automatically connected to the registered Crossref Funder DOI and is thus transparent, unique across systems and persistent. The connection between published work and funding agency is searchable in the Crossref Funder Registry⁴⁵ and also visible in the PID Graph.

```

"grants": [
  {
    "code": "777523",
    "links": {
      "self": "https://zenodo.org/api/grants/10.13039/501100000780::777523"
    },
    "title": "Connected Open Identifiers for Discovery, Access and Use of Research Resources",
    "acronym": "FREYA",
    "program": "H2020",
    "funder": {
      "doi": "10.13039/501100000780",
      "acronyms": [],
      "name": "European Commission",
      "links": {
        "self": "https://zenodo.org/api/funders/10.13039/501100000780"
      }
    }
  }
],

```

Figure 14 Integration of grant IDs and Crossref funder DOIs in the JSON metadata

InvenioRDM is going to be available as a beta release late 2020 and will include support of ROR organizational identifiers. The integration of ROR IDs was already achieved and will be part of the first release.⁴⁶ Having ROR IDs integrated in InvenioRDM contributes to a wider integration of ROR IDs across repositories, as the software already includes this feature and the effort to integrate them in a services becomes thus lower. For Zenodo, this will become visible once it switches to InvenioRDM, when users will have the opportunity to link their affiliations in future records to a ROR ID.

⁴⁵ <https://search.crossref.org/funding>

⁴⁶ <https://github.com/inveniosoftware/idutils/pull/59>

Lessons learned / Foreseen next steps

CERN Open Data portal

ORCID and ROR identifiers in the CERN Open Data portal can only be applied to a limited amount of records. Most outputs on the portal have a collaboration as the author (e.g. CMS collaboration which comprises thousands of authors) rather than individual authors, so for those cases it is not possible to assign individual ORCID iDs or ROR IDs. As of October 2020, the work on the new ORCID and ROR functionality is finished and available on the production system⁴⁷.

Zenodo

Zenodo will fully migrate to InvenioRDM until the end of 2020, including all PID related features of InvenioRDM. Zenodo will maintain its support of Crossref Funder IDs and grant IDs, but also support for ROR IDs by the end of 2020 in its live system. During the implementation process, it was not clear how the checksums of ROR IDs are calculated and how the validity of entered strings can be verified. ROR IDs are URLs that resolve to the organization's record and they are described by a "unique and opaque character string: leading 0 followed by 6 characters and a 2-digit checksum"⁴⁸. However, as it was not specified how exactly the checksum was calculated, it was not clear enough how ROR IDs can be validated when users enter values and thus if the entered string is correct.

⁴⁷ Example of a record with ROR and ORCID links: <http://opendata.cern.ch/record/463>

⁴⁸ <https://ror.org/facts/#core-components>

3.5 PANGAEA

PID resource selected

This work describes the expanded integration of **ROR PIDs** and **grant IDs** (*not PIDs*) in PANGAEA data publication metadata.

Community served

PANGAEA is a data publisher for earth and environmental research data serving a large international community of researchers. PANGAEA provides state of the art data publication services with one-on-one data archiving and publication support. It receives long term funding from the Alfred Wegener Institute in Bremerhaven and the MARUM-University of Bremen, while it offers its services to anyone in need of publishing data in a curated data archive for archiving, citation, and dissemination purposes. It works closely with its stakeholder community (individual researchers, research institutes, publishers and national data infrastructures) to adapt its services and workflows continuously to new standards and best practices and to best support open and FAIR science.

Recently, publishers have been pushing actively to enforce data publications as a mandatory part of publications of research results, which shows the very timely developments moved forward in the FREYA project. By enriching dataset metadata with relevant identifiers, in this case ROR IDs and Grant IDs, the impact of these aspects of the research effort can be better tracked and demonstrated as it links the various research components.

PANGAEA serves the larger natural science community, with a focus on the earth and environmental sciences. In addition, using the GFBio broker service⁴⁹, PANGAEA sustains a reciprocal dataset linkage with genetic data archived at ENA⁵⁰, thus providing a permanent (DOI) link between genetic resources and the environmental parameters relevant to the resource (for example see Fuchs *et al.* 2016⁵¹). Services provided by PANGAEA thereby also extend to researchers active in various fields of genetics. As a result of very recent efforts, PANGAEA is establishing protocols and workflows to include data from the social sciences as well.

Inclusion of ROR IDs and grant ID in dataset metadata will allow research institutes and funders to track the scientific output generated in terms of dataset publications and the attached downstream research outputs. PANGAEA is a curated data repository with a team of trained curation staff who, through their direct interaction with the users, also ensure that metadata is completed with the most relevant and correct information. Following the FREYA recommendation for the use of RORs to identify research organizations, PANGAEA staff responsible for the curation of data from the social sciences have already started to include the RORs in new dataset publications. The curation staff has been informed about the implementation of changes in regards to grant IDs and RORs and have been instructed to include this information whenever possible going forward, while the work reported here also focuses on adding this information to our current data holdings.

Specifications for Integration

Specifications for Grant IDs

The integration of grant identifiers focused on two dominant funding sources for research data in PANGAEA, the European Commission and the DFG (Deutsche Forschungsgemeinschaft). Unfortunately, grants awarded through the European Commission funding lines and the DFG are currently not issued persistent identifiers (grant DOIs). Although this may hopefully change in the near future, the current

⁴⁹ https://www.gfbio.org/de/gfbio_ev

⁵⁰ European Nucleotide Archive: <https://www.ebi.ac.uk/ena/browser/home>

⁵¹ Data publication by Fuchs *et al.* 2016: <https://doi.pangaea.de/10.1594/PANGAEA.860256>

situation forced us to focus on the integration of unique identifiers with landing pages provided by the respective funders, which we still considered to provide a benefit to the research community allowing users to navigate to grant descriptions and search for datasets related to grants in PANGAEA.

Specifications for ROR IDs

PANGAEA performed the following steps to determine the specifications for integration of ROR PIDs:

1. ROR PIDs had to be matched to the more than 9000 entries in the PANGAEA organizations database (PAN.DB) to allow the integration of RORs in the currently published dataset metadata.
2. The ROR API was used to gather RORs by matching with entries from PAN.DB using two different methods: 1) Querying & filtering 2) Affiliation matching, showing that Affiliation matching returned a much larger number of ROR matches (which still contained a large number of false positive and false negative errors). This resulted in having to manually check the API output and correcting for false negative and false positive results.
3. We had to assess the optimal places to include RORs, as changes in PAN.DB registry carry through to all fields that the registry is linked to (relational database architecture). It was decided to expand integration of ROR from organizations named in dataset publication titles to become an integral part of the metadata used to describe projects.
4. RORs have been developed to address the affiliation use case: “to unambiguously identify which organizations are affiliated with which research outputs”. The integration of RORs in different parts of the metadata had to accurately reflect the complexity of the research environment, where often several organizations contribute to the research in different ways. The use case is, however, a very general one and we focus most on populating the PID landscape with organizational identifiers so that any use case involving organizations as an endpoint or intermediate identifier can be realized via PID Graph APIs involving PANGAEA records (via Datacite or schema.org).

Integration achieved

Integration of Grant IDs in PANGAEA

Funder IDs (Crossref Funder Registry) are already an integral part of the dataset metadata published through PANGAEA and were now completed with the addition of IDs for grants, where this was possible. Figure 15 shows a published dataset by Stenvers *et al.* 2020⁵², which has received funding from both the EU Horizon 2020 program and the DFG, so is used here to show the placement and functionality of the grant identifier. So far, PANGAEA’s relational database cannot be searched for output related to individuals grants, however, a filter for “projects” related to grants can be applied for the search until project PIDs can be implemented. Users are redirected to the respective funder’s registry (EU funding – Cordis and DFG funding – GEPRIIS), allowing them to collect more information about the grant that funded the research.

⁵² Stenvers et al.: <https://doi.pangaea.de/10.1594/PANGAEA.918915>

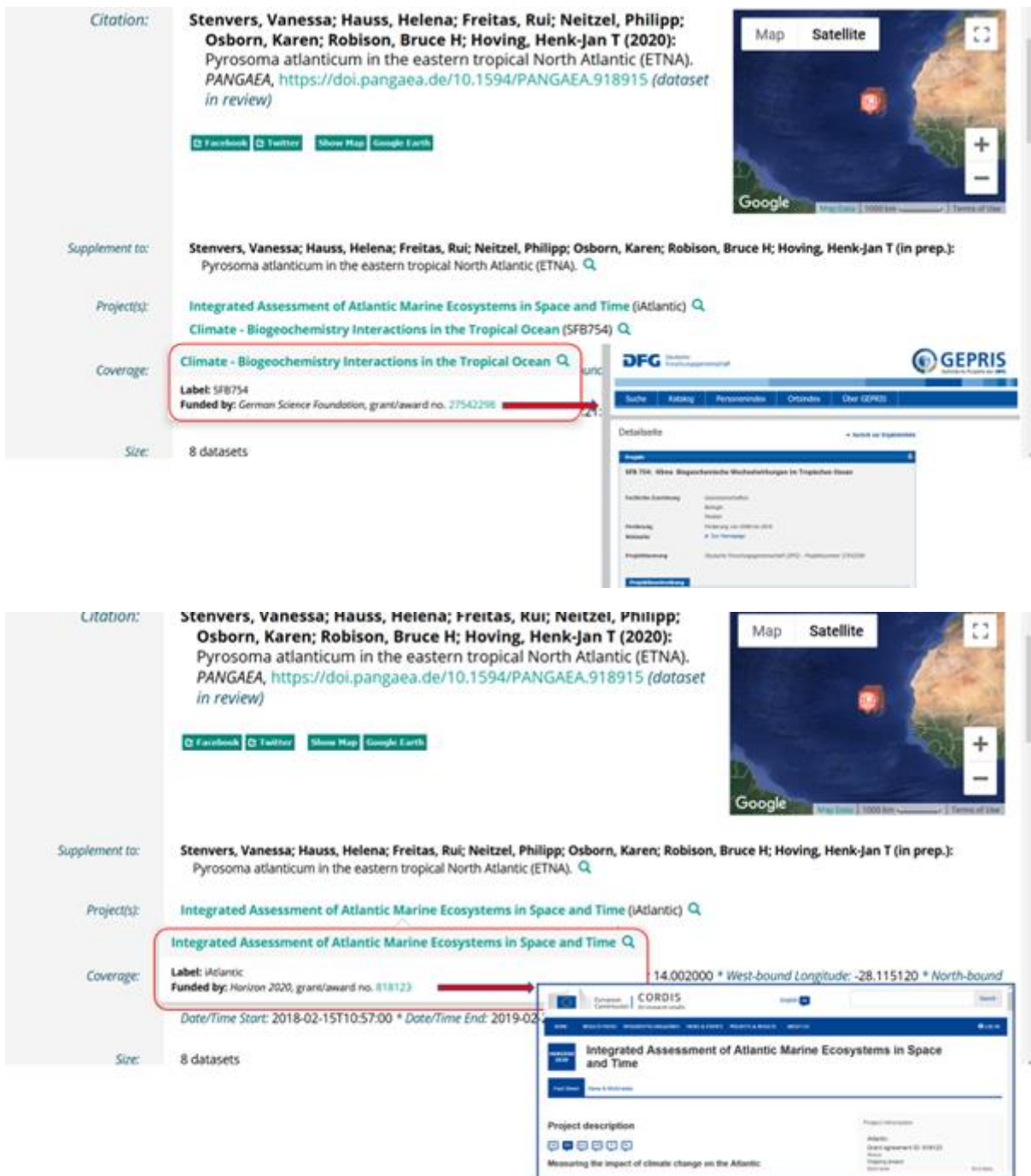


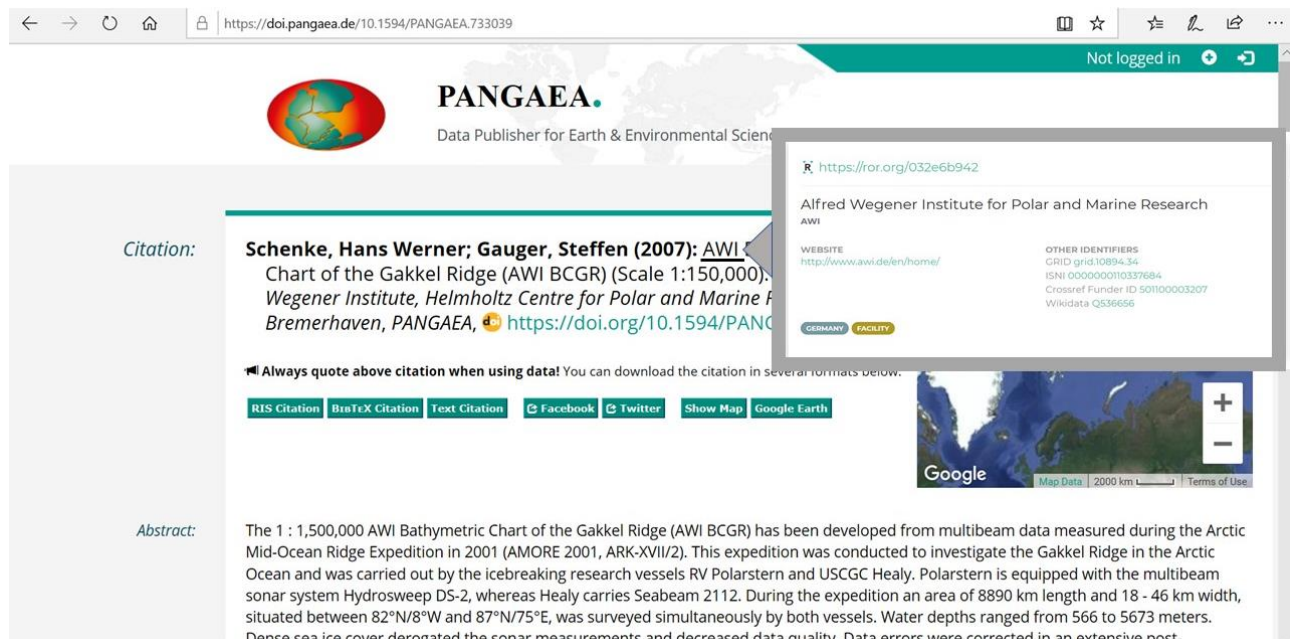
Figure 15 Dataset by Stenvers et al. 2020, now including actionable grant IDs for funding from the Horizon 2020 program and the DFG

The “Project” label pop up windows provide additional information like funder and actionable grant IDs and will also include an actionable ROR PID for the project coordinators affiliation. Grant IDs have been added to the metadata of existing records and will be used on all relevant datasets in future publications. As soon as grant PIDs become available for these and other funding lines, the grant IDs will be replaced by grant PIDs (DOIs).

Integration of ROR IDs in dataset metadata

The integration of PIDs for research organizations (RORs) in PANGAEA is still underway, but has been expanded from the original integration plan, which has been described in more detail in FREYA Deliverable 4.4. Efforts were initially only focused on integrating RORs in dataset titles, when organizational names

were part of the title (see Figure 16). In this case, the actionable ROR identifier pops up when hovered over, so that users could navigate to the institution's landing page in the ROR registry.



The screenshot shows a web browser displaying a PANGAEA dataset page. The URL is <https://doi.pangaea.de/10.1594/PANGAEA.733039>. The page features the PANGAEA logo and the text "Data Publisher for Earth & Environmental Science". The citation is: **Schenke, Hans Werner; Gauger, Steffen (2007): AWI Bathymetric Chart of the Gakkel Ridge (AWI BCGR) (Scale 1:150,000). Wegener Institute, Helmholtz Centre for Polar and Marine Research, Bremerhaven, PANGAEA, doi:https://doi.org/10.1594/PANGAEA.733039**. A pop-up window shows the ROR identifier <https://ror.org/032e6b942> for the Alfred Wegener Institute for Polar and Marine Research (AWI). Other identifiers listed include GRID grid:10894-34, ISSN: 0300-0001/0337684, Crossref Funder ID: 5011000033207, and Wikidata: Q536656. The abstract describes the development of a bathymetric chart from multibeam data during the Arctic Mid-Ocean Ridge Expedition in 2001 (AMORE 2001, ARK-XVII/2).

Figure 16 A PANGAEA dataset with a Research Organization included in the title of the data publication, in this case the Alfred Wegener Institut (AWI). The “AWI” in the title resolves to the record in the ROR registry once the “cleaned” ROR matching list is imported. Also visible in the ROR registry are other identifiers linked to the research organization.

Although the integration of RORs in other parts of the metadata where “institutions” are included (e.g. author affiliations) is wanted and sensible, the architecture of PANGAEA, founded on a relational database, restricts how RORs can be integrated, since an update to e.g. an author's affiliation, would change this information in other fields as well. In addition, the contribution of different research organizations to a single project can be very complex (owner of ship, coordinating organization, dataset author affiliation) and we are still working on finding sensible solutions to portray these relationships. The integration has, however, been expanded to include RORs in “project” label metadata to identify the organization that coordinated the project. This integration is achieved by including RORs for the project coordinator's organization, which is already part of the project metadata but without an organization PID. The project metadata relating to the coordinating institution will now also be visible and actionable to users via a pop-up window (see Figure 17) as soon as ROR identifiers are imported as semanticURI.

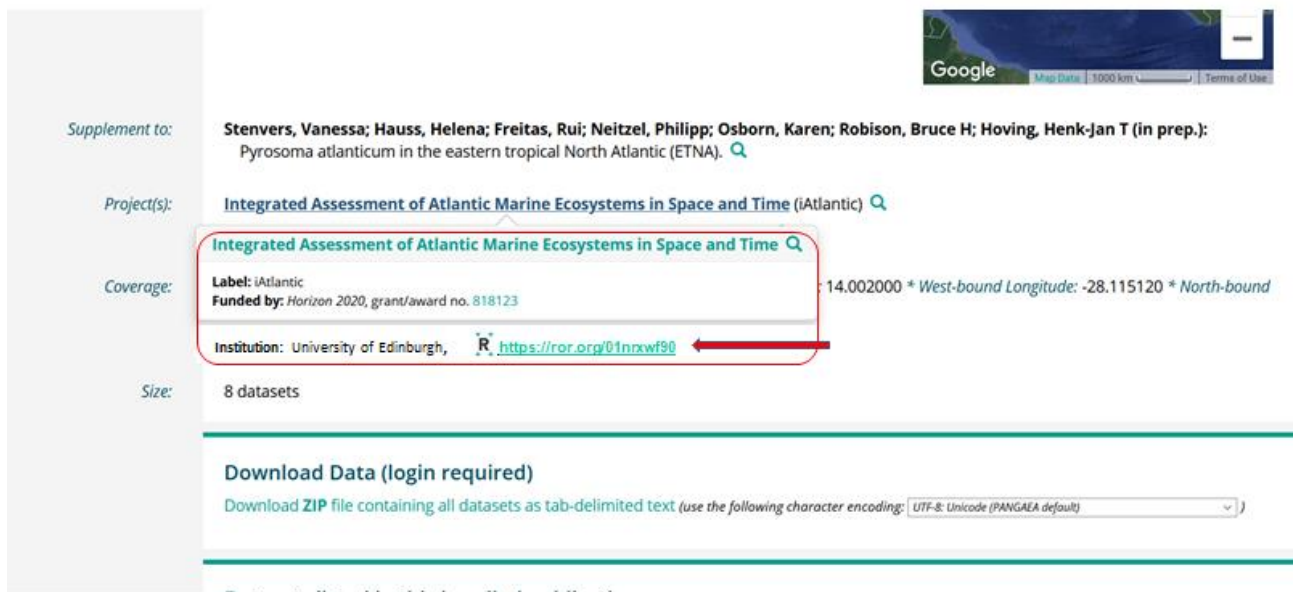


Figure 17 PANGAEA dataset publication example to include actionable ROR identifier in user facing project metadata on splash pages

PANGAEA will retain double entries (ROR and Crossref funder IDs) for organizations that function as both funders and project coordinators, since RORs are meant to apply solely to research organizations in the stricter sense. Depending on the organization’s role in relation to the published research, the appropriate identifier will be applied.

Work matching the PANGAEA organization registry to the ROR registry included analytical steps to understand the benefits and drawbacks of either approach (“querying and filtering” vs. “affiliation matching”) and are provided through a GitHub repository⁵³. Overall, “affiliation matching” provided a much higher number of retrieved RORs than “querying” and was deemed the better approach (Figure 18).

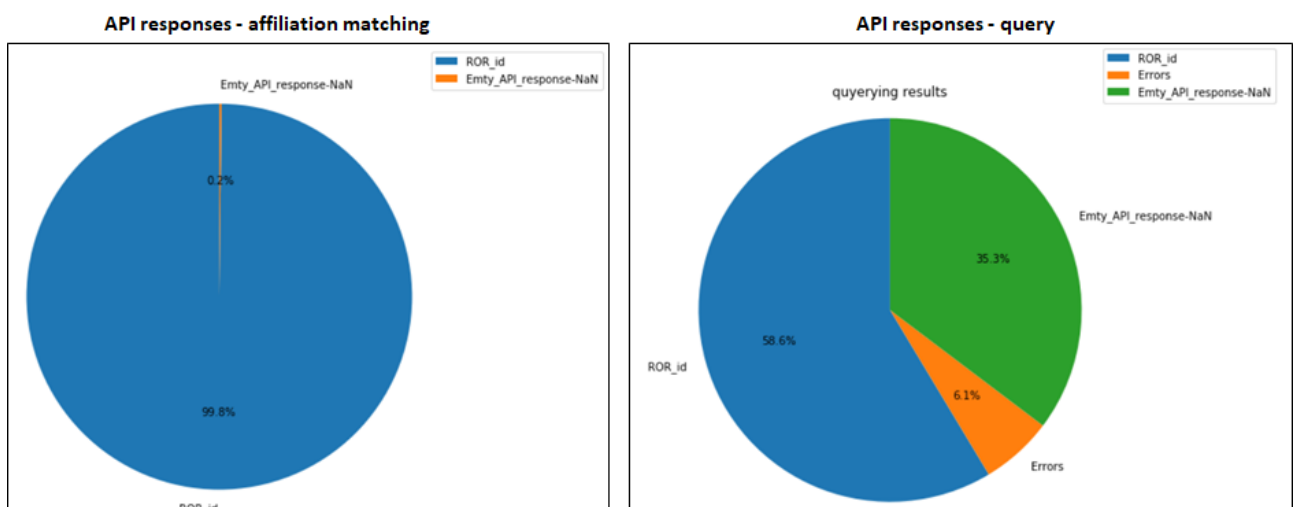


Figure 18 “Affiliation matching” returned significantly more ROR_id’s than the “query” method

⁵³ <https://github.com/pangaea-data-publisher/ROR-matcher/blob/b94ebabe448988c28f6b7aa71644cb835796d9a0/>

ROR API

“Querying” – results (Figure 19)

For 58.6% of Pan.DB entries some ROR_id was found (not always the correct one)

For 35.3% of Pan.DB entries nothing was found by the ROR API – empty response.

In 6.1% of Pan.DB entries have some error: KeyError, TypeError

Error types:

KeyError – API returns an error because some of the input fields were incorrect. For example “country”= 'New Jersey, USA’

TypeError – some of the input fields (country or name) are empty

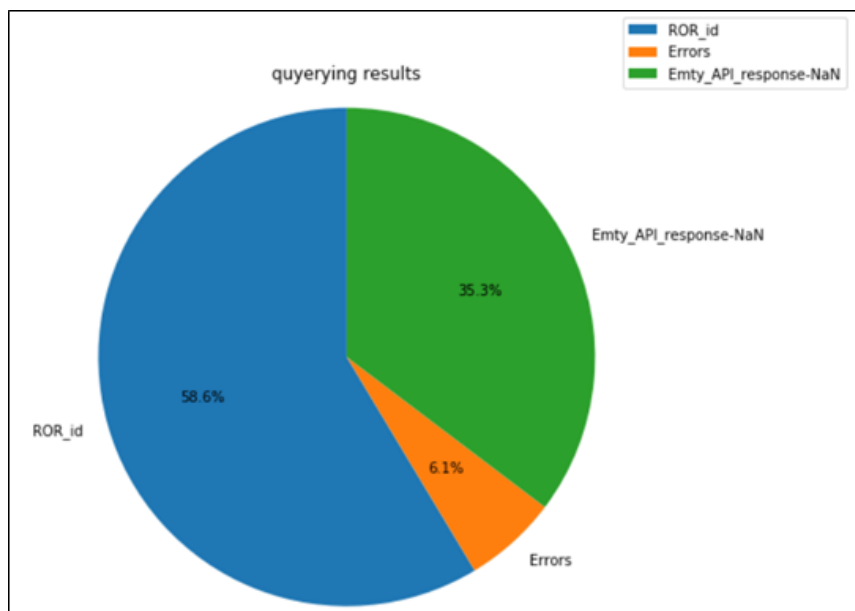


Figure 19 “Querying” results for PAN.DB and ROR registry matching

ROR API

“Affiliation matching” – results (Figure 20)

For 99.8% of Pan.DB entries some ROR_id was found (not always the correct one)

For 0.2% of Pan.DB entries nothing was found by the ROR API – empty response

56.7% of ROR_ids were labeled FALSE by API (some of these IDs are actually valid when checked manually)

43.3% of ROR_ids were labeled TRUE by API

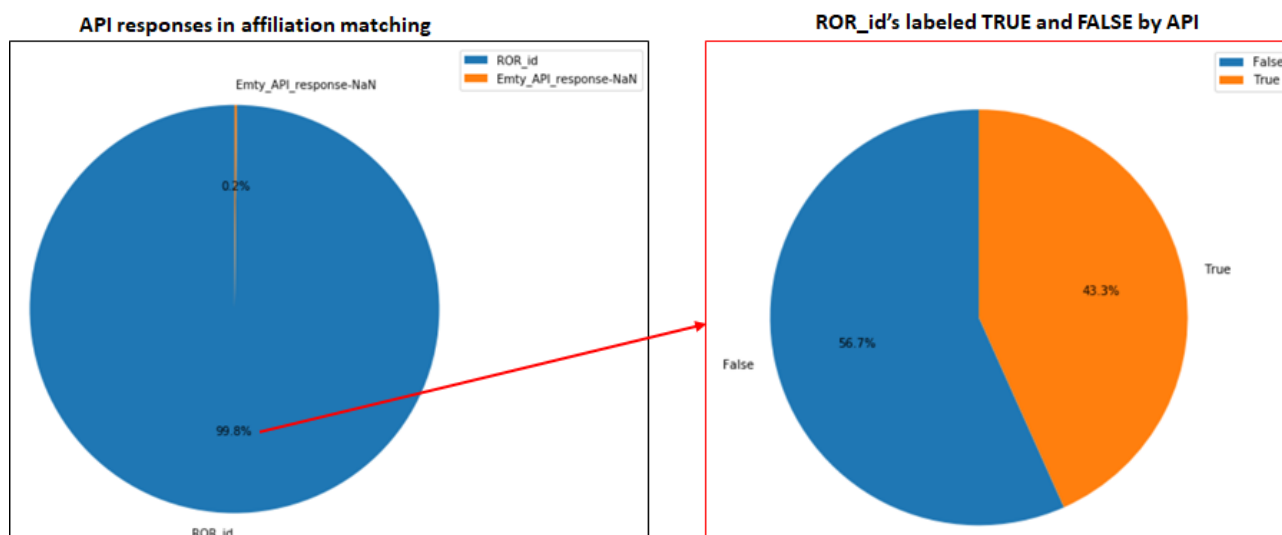


Figure 20 "Affiliation matching" results for PAN.DB and ROR registry matching with a more detailed analysis of API response true/false rating of matches found

Output from the ROR API was used as a basis to start manual checking of the > 9000 records in PAN.DB that returned potential ROR matches. API scoring of matches did not prove useful for filtering as criteria for scoring may be too strict, providing a large number of false negatives. However, false positives on high scoring entries were also a common error. Once manually checked, the completed list will be imported and included in the described metadata fields to enrich the dataset metadata with definite and persistent identifiers for research organizations.

Lessons learned

The approach chosen by PANGAEA required a lot of man hours since work for the API approach had to be repeated with manual checking, due to such high error rates in matching. Going forward, the inclusion of RORs will be much less problematic as the records are either already in our database or can be directly retrieved from the registry. However, investments in approaches like the one provided by EMBL-EBI in this deliverable are very sensible solutions and have shown to be applicable across institutions (STFC). The needed approaches rely heavily on the data structure of the registry to be matched and a suite of solutions targeting different matching strategies would be very useful and would provide great incentive for the retrospective implementation of RORs.

Within the ROR registry, there are inconsistencies in regard to the granularity of organizations listed. For example, the MARUM Department of the University of Bremen has its own ROR, even though RORs should not resolve at the level of university departments. These errors are still quite prevalent in the registry but will surely subside with increased use and feedback from the community. However, response rates to error reporting and improvement suggestions by us during the implementation phase have been very low and the ROR effort would benefit greatly from a more active engagement with the research community at the registry level to ensure that the gathered momentum can be sustained.

3.6 STFC

PID Graph resource selected

STFC focussed on using relevant PIDs in the prototype of the STFC Open Science Portal that is described in more detail in Deliverable 4.7. In addition to the established PID types such as DOIs for research papers and datasets, the following emerging and new PID types have been explored with the purpose of their integration in the Portal: ROR IDs and GRID IDs for organizations.

We explored other types of persistent identifiers and made preparatory work for their adoption in our research environment. This effort is outlined in the “Integration achieved” subsection.

Community served

The target community reflects the complex nature of STFC operation, as it is a funder of science and postgraduate education, also a research organization, with beneficiaries of synchrotron radiation beamtime in a range of research disciplines, from physics to biology and occasionally in humanities. STFC as a funder would be more interested in seeing the use of Funding PIDs, Project PIDs and PIDs for organizations – so one part of the community can be designated as Funding Managers; STFC as a research organization would appreciate more the use of PIDs for facility instruments and experimental samples – so another part of the community is Facility Instrument Scientists and Visitor Scientists. Comprehensive research impact studies may involve any of the aforementioned PID types, so another user category would be Research Impact Managers assisted by Research Information Managers.

Integration achieved

STFC Open Science Portal harvested records of science from various open sources and tried to enrich them where possible with the PIDs. Different levels of integration have been achieved for new and emerging PID types:

ROR IDs and GRID IDs for organizations have been added to organizations (mostly UK universities) funded by STFC for their research projects or postgraduate studentships, i.e. addition of IDs to existing records. This integration is going to be demonstrated in the STFC Open Science Portal prototype. In addition, we collaborated with EMBL-EBI who helped us to match DOIs of a few thousand publications in STFC repositories with the RORs of corresponding organizations (authors’ affiliations) using machine learning techniques.⁵⁴ Assigning PIDs for organizations allowed us to put facility experiments in a richer information context and clearly visualize which organizations have been involved, what are their roles and what are relations between organizations. This is illustrated by Figure 21.

⁵⁴ See section 2.1 for a presentation of Europe PMC’s ML mapping results
https://docs.google.com/presentation/d/1fqQq4dISwJN0T2L42LI2Wfb9lfcSuxg_LicQHT1b7-Q/edit?usp=sharing

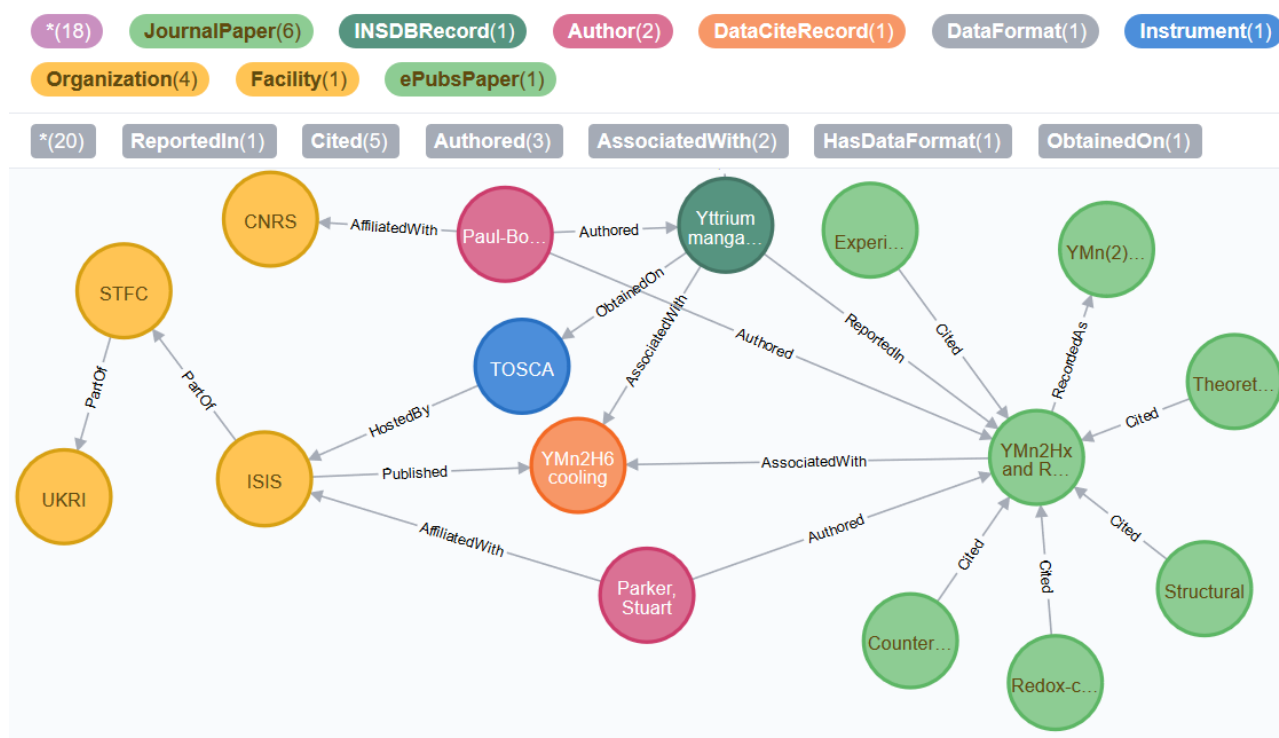


Figure 21 Rich research information context around INSDB (Inelastic Neutron Scattering Database) record that involves a description of the original experiment represented by the DataCite record, people and organizations involved, as well as the resulting research paper cited by other papers. There are different types of links for organizations, representing their roles in this piece of research and their mutual relations.

PIDs for facility instruments have been explored in depth, both technical/metadata aspects of their implementation and new practices required for their genuine adoption by STFC facilities. DataCite metadata profiles following the RDA Instrument PIDs recommendations have been produced for STFC facilities, and discussions are ongoing with facility instrument scientists about adoption of instrument (beamline) PIDs. STFC Open Science Portal has done much for the facility instruments disambiguation (where instruments are still identified by their unique commonly known names), and this should be seen as preparatory work for the further adoption of the beamlines PIDs. Challenges for the sustainable adoption of the facility instruments PIDs are reiterated in the “Lessons learned” subsection. Figure 22 shows a DOI metadata template for TOSCA instrument on ISIS neutron and muon facility. We are engaging with the EXPANDS project⁵⁵ who have a task for PIDs in facilities, trying to get EXPANDS interested in the actual adoption of these metadata profiles.

⁵⁵ The European Open Science Cloud (EOSC) Photon and Neutron Data Service (ExPaNDS) project. <https://expands.eu/>

```

<?xml version="1.0" encoding="UTF-8"?>
- <resource xsi:schemaLocation="http://datacite.org/schema/kernel-4
  http://schema.datacite.org/meta/kernel-4/metadata.xsd"
  xmlns="http://datacite.org/schema/kernel-4" xmlns:xsi="http://www.w3.org/2001/XMLSchema-
  instance">
  <identifier identifierType="DOI">????/??????</identifier>
  - <creators>
    - <creator>
      <creatorName nameType="Organizational">ISIS Neutron and Muon Source</creatorName>
    </creator>
  </creators>
  - <titles>
    <title titleType="Other">TOSCA indirect geometry spectrometer</title>
  </titles>
  <publisher>STFC UKRI</publisher>
  <publicationYear>2020</publicationYear>
  <resourceType resourceTypeGeneral="Other">Instrument</resourceType>
  - <relatedIdentifiers>
    <relatedIdentifier relationType="IsDescribedBy"
      relatedIdentifierType="URL">https://www.isis.stfc.ac.uk/Pages/Tosca.aspx</relatedIdentifier>
    <relatedIdentifier relationType="IsDescribedBy"
      relatedIdentifierType="URL">https://www.isis.stfc.ac.uk/Pages/TOSCA-
      History.aspx</relatedIdentifier>
  </relatedIdentifiers>
</resource>

```

Figure 22 A metadata template prepared for a facility instrument according to the RDA PIDINST WG recommendations using DataCite schema

PIDs for experimental samples have been considered but their introduction is postponed owing to the actual practices of visitor scientists who bring their samples onto facilities. Openly published information about these samples is often deliberately generic, up to the point of obscurity, as the nature of the samples exposed on the beamline can alarm research competitors if the information about samples is openly published. It is unlikely that PIDs for samples in research facilities with any sort of sensible metadata associated with the PIDs can be introduced soon.

PIDs for software are de-facto minted by a few STFC teams, with the main publishing venues for them being DataCite and Zenodo. A smaller demonstrator with a few dozen software records identified by their PIDs (that include PIDs for software versions) is on the way to the STFC Open Science Prototype. We looked in depth in the case of MANTID software package for experimental data analysis⁵⁶ and the use of various ways of MANTID citations over the 5-year period. Despite having DataCite DOIs for the master record of MANTID and for each particular version released, there is a strong tendency towards citing the associated journal paper⁵⁷ rather than the DataCite records. Another observation concerns the attitudes of some publishers who do not propagate citations of DataCite records down the research information value chain, e.g. a publisher may supply CrossRef with references to journal articles but exclude DataCite records from the list of references supplied.

The aggregation of software citations obtained through citing a DOI of a journal article and DataCite DOIs was fed into the disciplinary analysis of the software use represented by Figure 23, yet with the current practices of software citations, it looks like it only makes sense to count citations of the corresponding journal article.

⁵⁶ Mantid (2013): Manipulation and Analysis Toolkit for Instrument Data.; Mantid Project.

<http://dx.doi.org/10.5286/SOFTWARE/MANTID>

⁵⁷ O. Arnold, et al., Mantid—Data analysis and visualization package for neutron scattering and μ SR experiments, Nuclear Instruments and Methods in Physics Research Section A, Volume 764, 11 November 2014, Pages 156-166, <http://dx.doi.org/10.1016/j.nima.2014.07.029>

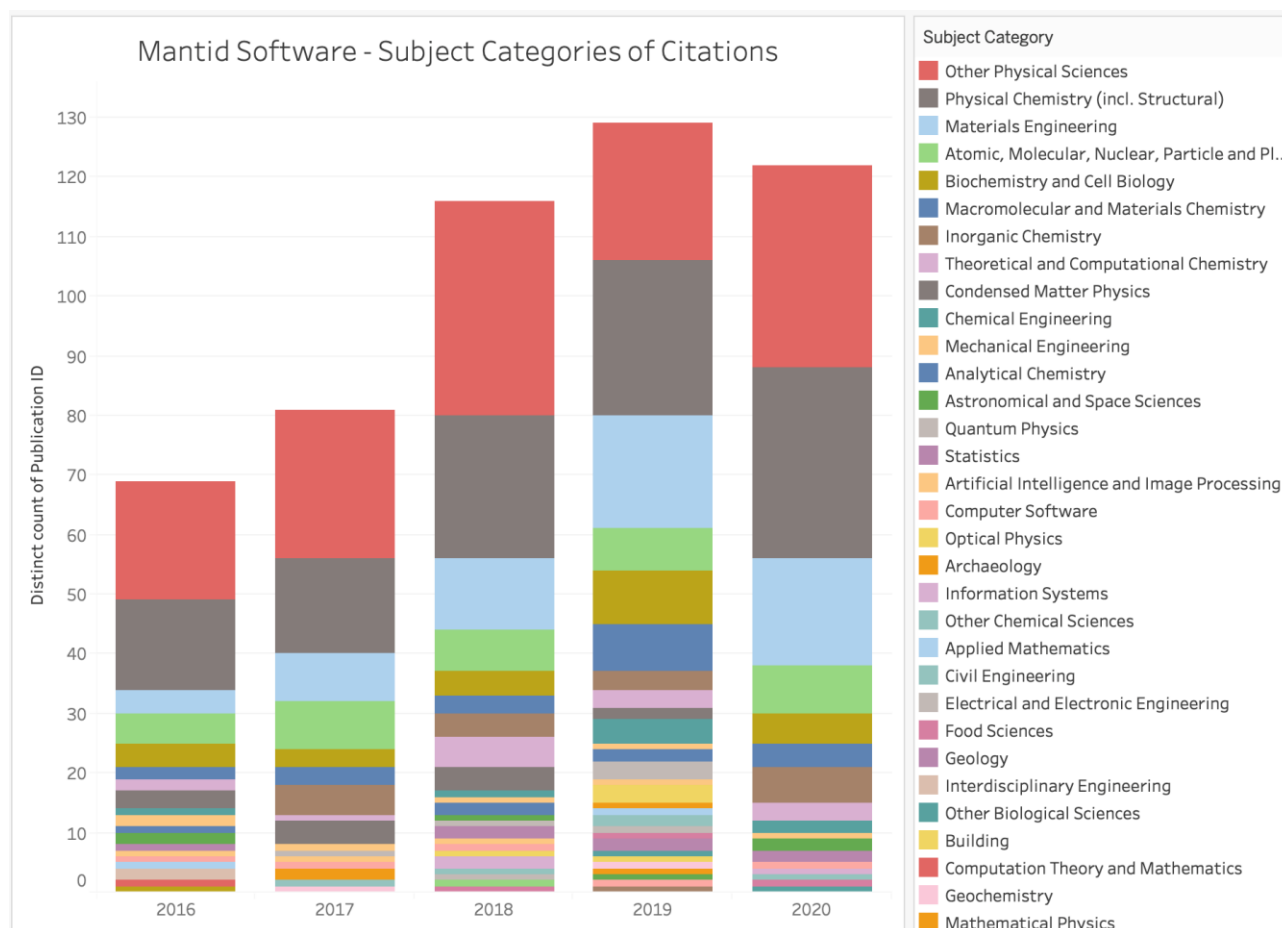


Figure 23 Disciplinary breakdown of MANTID software citations

Funders PIDs have been harvested from the GtR (Gateway to Research)⁵⁸ – the main UK source of research funding public information. These PIDs are unique within the GtR and cannot be considered fully-fledged PIDs, e.g. with the purpose of their citation. The potential uptake of the funders PIDs by STFC and other UK Research Councils is unclear at the moment, yet the ongoing restructuring of the UK research landscape when all Research Councils are being amalgamated in a single UK Research and Innovation body may present an opportunity to introduce new sensible practices for funding identification. These opportunities will be explored after the FREYA end and will be driven by the implementation and demonstration needs of the STFC Open Science Portal.

Project PIDs, similarly to Funding PIDs, have been harvested by the GtR and integrated in the STFC Open Science Portal prototype. The issues and opportunities for these Project PIDs are the same as for the Funders PIDs.

Lessons learned / Foreseen next steps

Technical aspects of the new and emerging PID types adoption have proved to be a much lesser challenge than actual practices of the stakeholders involved. The main problem with the practices can be described as the problem of authority: who should be the authority for a new PID type in case when such authority does not exist? As an example, ROR identifiers for organizations do have the authority of the ROR consortium behind them, so RORs integration in STFC Open Science Portal or any other information system is just a technical matter. For research instruments though, there is no similar well-established authority, and instrument owners may not be motivated or committed enough to sustain the new kind of PIDs. There is an

⁵⁸ Gateway to Research <https://gtr.ukri.org/>

ongoing discussion what can be done in situations like this; one possible solution is that the research library or research archive can take the responsibility for certain new PID types, yet they should be seen as an authority by respective stakeholders, and commit themselves to the long-time effort of maintaining the new PIDs and PIDs metadata.

Another observation that could be of a common interest is about formal citations of PIDs-assigned software, stakeholders do not propagate these citations down the research information value chain, this has to be tackled through sensible discussions in suitable forums that could persuade publishers to change these attitudes.

FREYA managed to raise awareness of STFC stakeholders about new PID types, and made preparatory work in some cases with metadata elements disambiguated and well-prepared for the further PIDs adoption – as again in the case of facility instruments. This work will continue within STFC after the FREYA project ends.

4 Lessons learned and concluding thoughts

The implementations reported in the preceding chapter, presented FREYA partners with different challenges, yet all are indicative of the newness of the PID infrastructures.

The identifiers that were implemented included:

- ROR IDs by FREYA partners EMBL-EBI, the British Library, CERN, PANGAEA and STFC;
- Additional organization identifiers, GRID IDs and Wikidata IDs, by the British Library and STFC;
- Identifiers for grants (those that are internally assigned by a funder, not global grant DOIs) by DANS, CERN, PANGAEA and STFC;
- Identifiers for Funders (Crossref funder IDs) by the British library, CERN and PANGAEA;
- ORCID iDs by CERN (although these belong to a mature PID infrastructure - there are some lessons that linger).

The implementations involved adding identifiers to existing database records, retrospectively or providing the ability for identifiers to be incorporated into any newly added records. Where provision was made for future records to include new identifiers, a field was generated in the metadata to accommodate the identifier and relevant steps added to a user workflow for submitters to include the identifier.

There are some general lessons that could be useful for EOSC stakeholders, unfunded partners and others in the community wishing to undertake similar implementations:

Mapping affiliations to ROR IDs is difficult

ROR IDs were added retrospectively to records in Europe PMC, CERN's Open Data portal, PANGAEA's published dataset metadata, and STFC's soon-to-be-launched Open Science Portal. A key challenge was to map organization names within existing records to ROR IDs: the proportion of matches obtained with the ROR API, required enhancing by manual checking or by machine learning efforts. Since the ROR infrastructure is < 2 years old, the ROR registry, granularity and API are still evolving, and the capacity to address deficiencies is currently limited. This would be true for any emerging PID infrastructure.

Disparate types of identifiers being used for records

An example is given by DANS who found that not all stakeholder organizations use the same identifier types for grants or funding organizations. This has implications for cross talk between different organizations/repositories. The issue is being addressed to some extent by repositories holding a list of alternative identifiers for a record. For example ROR.org provides a list of alternative identifiers that are assigned to any particular organization (GRID, ISNI, Crossref Funder Registry, Wikidata). Likewise Europe PMC's grantfinder registry contains both the funder's internal grant identifier and the global grant identifier (DOI) where in use. This extends to other PID resources such as publications: Europe PMC includes the DOI, PMID and PMC ID for each publication where present. Encouraging repositories to include alternative identifiers where possible will aid mapping efforts and cross-talk.

Existing metadata schema and guidelines for emerging PID resources still require maturation

For this deliverable the British Library has reported on work that allows additional organization PIDs to be applied to new records coming into the British Library's Shared Research Repository. Although there is an agreed metadata schema for these organization identifiers, further clarification is required for some identifiers for example to aid formatting. The British Library also felt it would be useful to have guidelines

to consult for website display of identifiers on records. These are currently available from some PID providers for maturer PID resource types such as publications and researchers: eg ORCID⁵⁹, Crossref⁶⁰.

Guidance when records contain collaborative groups or consortia listed as contributors

CERN has reported a pilot project where organization identifiers are assigned to existing records in CERN's Open Data Portal. Implementation became challenging for records listing collaborations as there is no guidance currently on how to add organization identifiers to contributors listed as a consortium. While this might be mitigated to an extent by defining guidelines and additional information that can be validated by submitters for new records, ROR IDs may not be applicable to all of CERN's use cases currently. Handling of consortia in records is not a new problem - publishers have long faced the problem of defining "authorship" of a record and provide guidance regarding the information that should be submitted eg the International Committee of Medical Journal Editors (ICMJE) guidelines contain specific information about how to handle group authorship⁶¹. Moreover, the guidelines that exist are constantly contested, an example of which is the long standing debate about what constitutes a contribution to scholarly work. Groups like CASRAI have come up with CRediT, a taxonomy for Contributor Roles that lists 14 recognised contributor roles on scholarly outputs, amongst which is funding acquisition⁶².

When including ROR IDs in PANGAEA records, the team had to identify an approach where the ROR identifier would accurately reflect the contribution by the respective organization. The contributions to published research data are often very complex and this needed to be considered in the approach. Therefore, complex contributions to existing records were handled as follows: only the organization of the project coordinator is linked to its ROR record as this role is clearly defined and has very few exceptions (i.e. shared project coordination between organizations). Often organizations can also fill the role of research funder. To differentiate the roles for the respective research publication, different identifiers are placed in published dataset metadata: ROR IDs are assigned for the project coordinator's organization, while the same organization acting as funder is assigned its Crossref funder ID.

The newer the PID infrastructure or concept of a PID resource, the greater the challenges

A big challenge with new initiatives is the incredibly slow uptake by communities because they do not want something that is not established and may not have a long life. Chapter 2.2 summarises some of the newer PID resource types, for which infrastructures are still being formulated, e.g. facility identifiers that have been pursued by STFC, and instrument identifiers pursued by PANGAEA. The lessons articulated by FREYA partners around these PID resources would apply to most new PID resource types. For instance, The need for a community-led governing body or other authority behind a new/emerging PID type: this would go some way to realising metadata requirements to be associated with a new PID resource. Such an authority would at minimum represent a commitment to sustain the identifiers and their metadata and more broadly a commitment to take responsibility for strategic decision-making about governance, and overall direction of development of the infrastructure.

The implementations here have aptly demonstrated that the ROR infrastructure, although clearly emerging and evolving, is more advanced than the global grant identifier system. Moreover, global grant identifiers have not been implemented by FREYA partners—this drives home the fledgling status of global grant identifier system that is only beginning to be adopted by funders. Implementing local grant IDs (those that have been assigned independently by each funder), despite their shortfalls⁶³, represents a step forward over a situation of not having grant identifiers at all. However, the FREYA partners urge all funders to

⁵⁹ Guidelines on the display of ORCID iDs in publications: <https://orcid.org/content/journal-article-display-guidelines>

⁶⁰ DOI display guidelines: <https://www.crossref.org/education/metadata/persistent-identifiers/doi-display-guidelines/>

⁶¹ ICMJE guidelines on authorship: <http://www.icmje.org/recommendations/browse/roles-and-responsibilities/defining-the-role-of-authors-and-contributors.html#two>

⁶² CASRAI and CRediT taxonomy: <https://casrai.org/credit/>

⁶³ <https://www.crossref.org/blog/wellcome-explains-the-benefits-of-developing-an-open-and-global-grant-identifier/>

consider using global grant identifiers (DOIs for grants) - a consistent system would reduce the need for mapping local IDs.

Emerging PIDs can be implemented for any discipline

Closing on a positive note, it is important to point out that the implementations described in this report provide models that are potentially far-reaching. As stated in the introduction, each reporting partner broadly represents a discipline, yet their institutional communities are both more nuanced *and* can be multidisciplinary. As an example of an integration serving multiple disciplines and then also a single research institution: Europe PMC based at EMBL-EBI, is a literature database primarily for life science researchers, yet the ROR ID integration reported is being used by others indexing research publications (such as FREYA partner, STFC, serving the facilities-based science community); the integration also addresses a very specific use case for EMBL-EBI - an institution wanting to determine its annual publication outputs. Different approaches can be used to map identifiers to existing records: PANGAEA used a different approach to EMBL-EBI to map ROR IDs to existing affiliation records: rather than a machine learning approach, they used the ROR API plus manual checking. PANGAEA's integrations stand to benefit multidisciplinary communities: PANGAEA focuses on indexing published datasets for the earth and environmental research, yet has begun to expand to the larger natural science community (notably genetics research) and even to social sciences. Furthermore, DANS, the British Library and CERN report on implementations to services that are offered to multidisciplinary communities in the Netherlands, UK, and globally to the High-Energy Physics community and Zenodo users, respectively.

Embracing nascent infrastructures and uncovering technical weaknesses is a crucial part of growing the infrastructure. In doing so, organizations such as those represented by FREYA partners provide a valuable service to the PID community.