DataCite
FIND, ACCESS, AND REUSE DATA

Harnessing the Power of the PID Graph

Robin Dasler, DataCite Product Manager
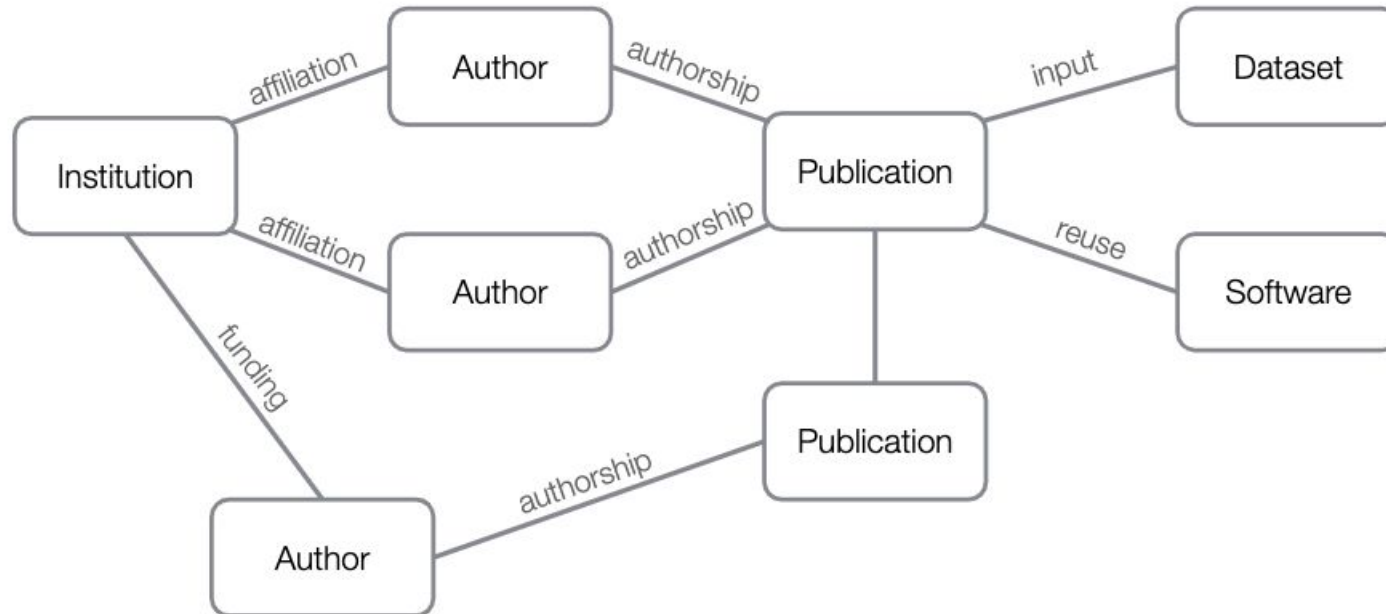
What is the PID Graph?

Research is already a graph

Researchers, institutions, publications, datasets, and more are interconnected.

Entities and the relationships between them form a conceptual graph of the connected research landscape.

PIDs are the backbone of connected research

Having unique persistent identifiers for researchers and their outputs is crucial to connecting pieces of the research landscape together.

PIDs already have the potential to enable the connected research graph, but we're not yet taking full advantage of their connecting powers.
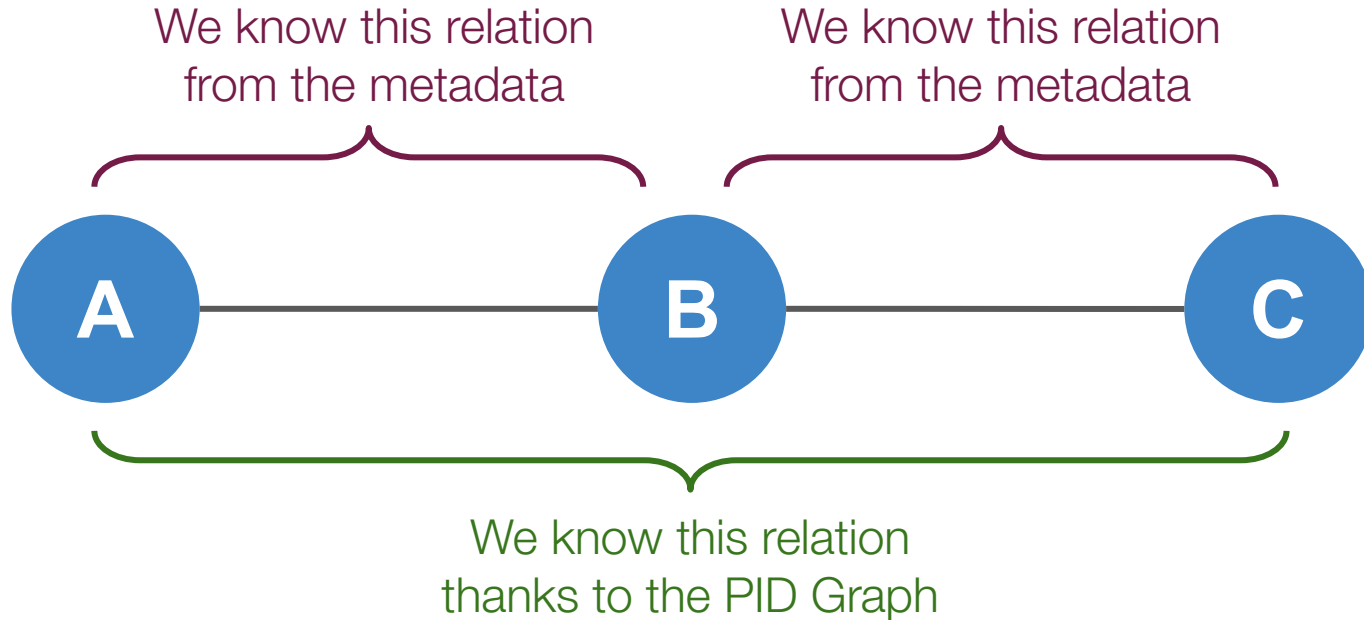
The **PID Graph**
(part of the EC-funded FREYA project) seeks to bring all these PIDs together into a connected graph.

The PID Graph concept

We can link PIDs together via relations in their metadata to enable the discovery of connections at multiple "hops" away.

We know this relation
from the metadata

We know this relation
from the metadata

A   B   C

We know this relation
thanks to the PID Graph

Working with the PID Graph

DataCite GraphQL API

The DataCite GraphQL API is currently in pre-release. It's ready for early adopters, but we're still expanding it and adding new features.

The API searches the metadata deposited with DataCite DOIs, as well as (currently) about 1.5 million related ORCID iDs and 3.5 million related Crossref DOIs. Adding information from ORCID and Crossref is an ongoing process.

Jupyter Notebooks

Jupyter notebooks are an open-source web app for building interactive documents that combine blocks of text with blocks of executable code.

DataCite has a GitHub repository for Jupyter notebooks where you can try out the GraphQL API. Anyone may fork the repo or submit their own notebooks via pull request.

https://github.com/datacite/notebooks

We'd love feedback on the GraphQL API no matter how you use it.

Let us know at https://www.pidforum.org/c/pid-graph

Live Demo

This notebook is a demonstration of the DataCite GraphQL API prepared for the midterm review of the EU-funded project FREYA. The notebook retrieves the datasets associated with the FREYA project, as well as all of the connections to those datasets, whether people or research outputs.

Note: At the time of notebook creation (June 2019), the DataCite GraphQL API is in pre-release and is not yet feature complete. It does not yet retrieve the metadata for `relatedIdentifiers` so for the purposes of demonstration we'll be referring to all of those items generically as "publications" because their actual resource type is unknown.

This notebook uses R, but you can use your language of choice in your own notebook. This notebook and others live in the DataCite Jupyter notebooks GitHub repository. Feel free to fork the repo and/or submit your own notebooks using the DataCite GraphQL API via a pull request.

Prepare the R GraphQL client

Load necessary libraries and set up the API endpoint.

```
library("httr")
library("ghql")
library("jsonlite")
library("IRdisplay")
library("dplyr")
library("igraph")

cli <- GraphqlClient$new(
  url = "https://api.datacite.org/graphql"
)
qry <- Query$new()
```

Generate the GraphQL query

In this query, we are looking through all DataCite DOIs that are assigned to datasets and finding those that include the FREYA grant number within a funding reference.

Then for each of those datasets, we're asking for:

1. Identifiers for the creators (in this case ORCID IDs)
2. Identifiers for items related to that dataset (citations, versions, etc.)
3. Identifiers for any funders related to that dataset

In this example, we already know that the datasets will have the EC as a related funder, but putting in the `funderIdentifier` will allow us to plot that information as part of our graph.

dataset
funder
publication
researcher
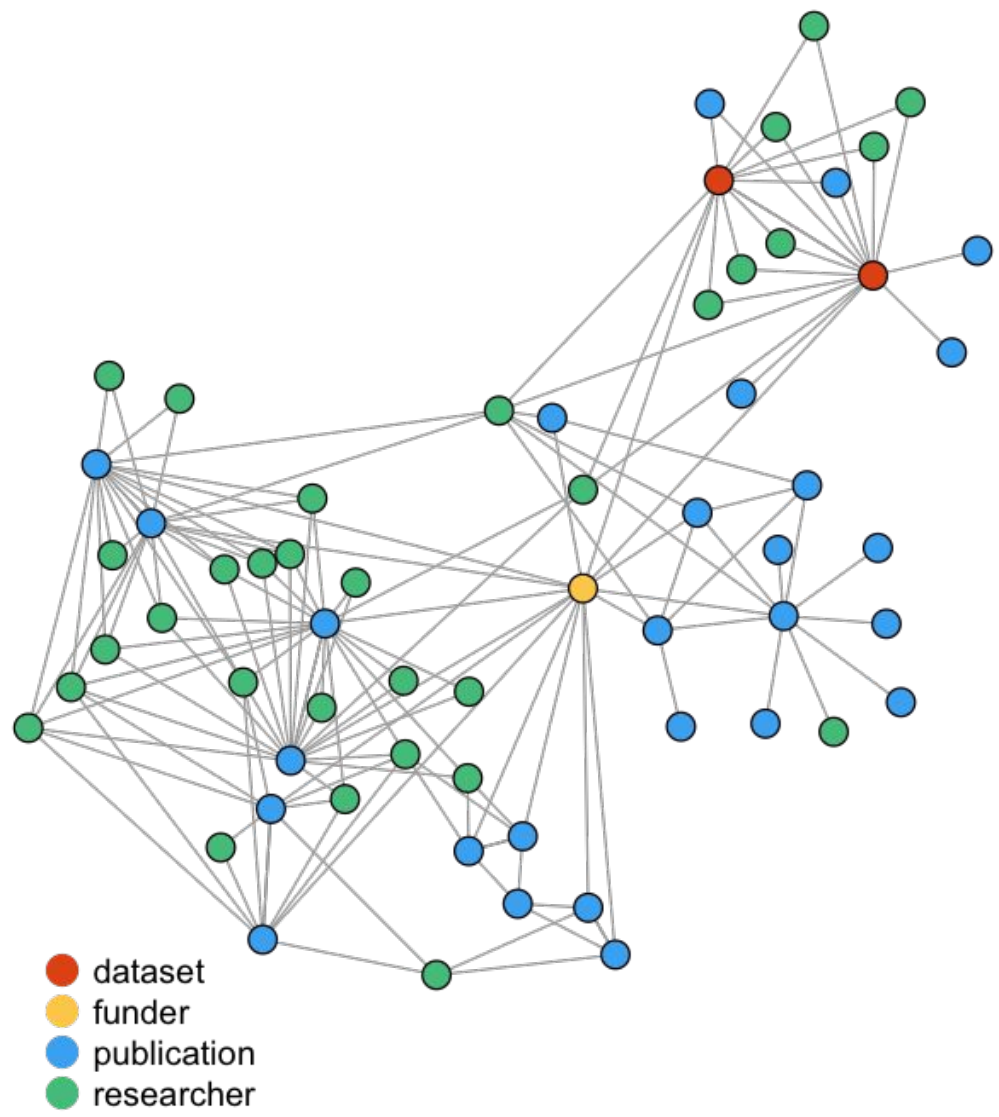
GraphQL Endpoint   https://api.datacite.org/graphql   Method   POST ⬍   **Edit HTTP Headers**

Graph*i*QL   (▶)   Prettify   History   ‹ Docs
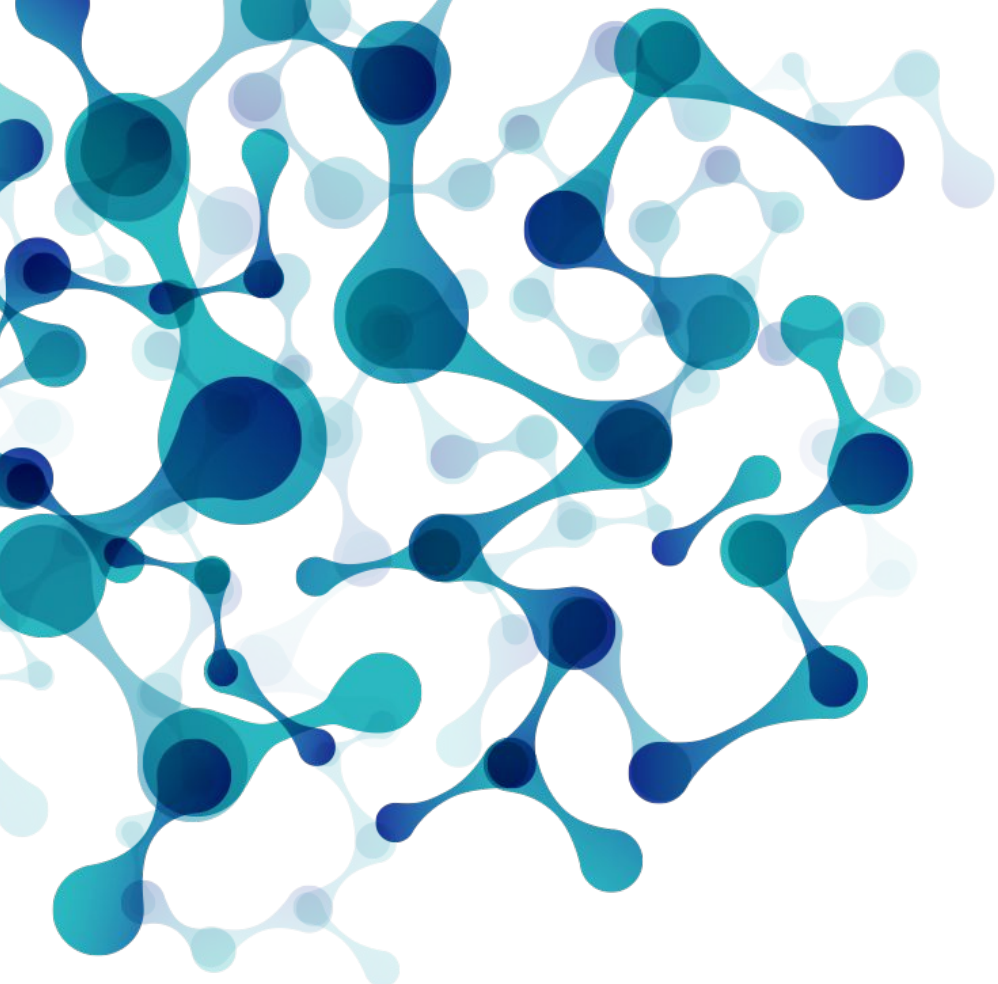
```
 1 ▾ {
 2 ▾   publications(query: "creators.name:dasler") {
 3       id
 4       titles {
 5         title
 6       }
 7       descriptions {
 8         description
 9       }
10 ▾     creators {
11         id
12         name
13         familyName
14       }
15 ▾     fundingReferences {
16         funderIdentifier
17         funderName
18         awardTitle
19         awardNumber
20       }
21     }
22   }
23
```

```json
{
  "data": {
    "publications": [
      {
        "id": "https://doi.org/10.5281/zenodo.1064000",
        "titles": [
          {
            "title": "Pid Service Adoption"
          }
        ],
        "descriptions": [
          {
            "description": "This presentation describes how
the uptake of persistent identifiers can be measured and gives
an overview of the main results of the ORCID adoption study."
          }
        ],
        "creators": [
          {
            "id": null,
            "name": "Dasler, Robin",
            "familyName": "Dasler"
          }
```

Thank you!

https://datacite.org

Twitter: @datacite