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

The PID Graph in FREYA
(additional project report)

Deliverable type	Report from Work Package 2 (not a contractual deliverable)
Dissemination level	Public
Due date	—
Authors	Martin Fenner (DataCite)
Abstract	The report describes work leading to the technical bases of the PID Graph in FREYA.
Status	Released 2 June 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.

Executive summary

This report describes the work done by FREYA work package 2 (WP2) on the PID Graph, one of the pillars of the FREYA project. It summarizes the early work on identifying user stories, describes the decisions about which technical architecture to implement, the work on launching a pre-release version of PID Graph core infrastructure in May 2019 in the form of a GraphQL API, and the work on refining this API until the launch of a production grade version in May 2020. This work has not been reported in detail elsewhere.

We then report on the integration roadmap going forward, with deliverables by FREYA WP2 and WP4 building on top of this core infrastructure, integrating services by core PID providers, FREYA disciplinary partners, and other EOSC partners into the PID Graph. FREYA WP5 (Iterative Engagement) has started work on broader community engagement and feedback collection via webinars, in-person workshops and FREYA ambassadors in May 2020, and puts a strong emphasis on Jupyter notebooks to integrate with the GraphQL API developed by WP2 to explore the PID Graph. The responsibilities for WP2, WP4, and also WP5 are described in the FREYA Description of the Action. No additional resources will be needed for this work.

By the end of the FREYA project in November 2020 the PID Graph will be supported by TRL-8 infrastructure that was first launched in May 2019, will be integrated in PID provider infrastructures via WP2, disciplinary infrastructures via WP4, other EOSC services via WP4, and the broader community via WP5. The GraphQL API supporting the PID Graph will be maintained beyond the duration of the FREYA project by DataCite and other PID providers, and will have become part of the common PID infrastructure.

Contents

1 What is the PID Graph?

Persistent identifiers (PIDs) are not only important to uniquely identify research outputs such as publications, datasets, or software, but the metadata for these persistent identifiers can also provide unambiguous linking to other resources, both other publications, datasets or software, but also actors such as researchers, research organizations, or funders. The FREYA project calls these networks of connected PIDs the *PID Graph*, and envisages new services operating over the Graph providing value to a wide range of users. Indeed the PID Graph is one of three motivating forces of FREYA, along with the PID Forum (a stakeholder community) and PID Commons (concerned with sustainability). However, the PID Graph itself is not a specific deliverable of FREYA, but rather underlies the project's technical work, and for this reason it is desirable to produce this additional report giving a concise account of what the PID Graph is in concrete terms and relating it to other developments in the project.

Establishing and taking advantage of these connections between resources identified by PIDs is difficult, and work is needed to connect existing persistent identifiers to each other in standardized ways. This will allow us to address important use cases, some of which are difficult or impractical to work on using the existing scholarly infrastructure.

Starting with user stories that describe important use cases, the FREYA project has built common infrastructure to address these user stories and is also working on specific client applications for these user stories.

2 User stories motivating the PID Graph

Starting in spring 2018, shortly after the beginning of the project, FREYA partners started to identify and collect PID Graph User Stories motivated by real needs of the partners' own organisations and communities. We collected them in a central place, the GitHub issues labelled "PID Graph"[1] in a code repository managed by DataCite. We collectively produced a total of 49 user stories related to PID Graph.

FREYA partners met in person on 22 August 2018, in London to add additional user stories, and discuss, group, and prioritize these user stories. We identified some user stories that did not require PID Graph functionality and others that seemed unrealistic to address within the time available for the FREYA project. But many use cases were highly relevant, and realistic to address. In the grouping and prioritization exercise we identified these three overarching themes that most (but not all) user stories fell into, as shown in Figure 1.

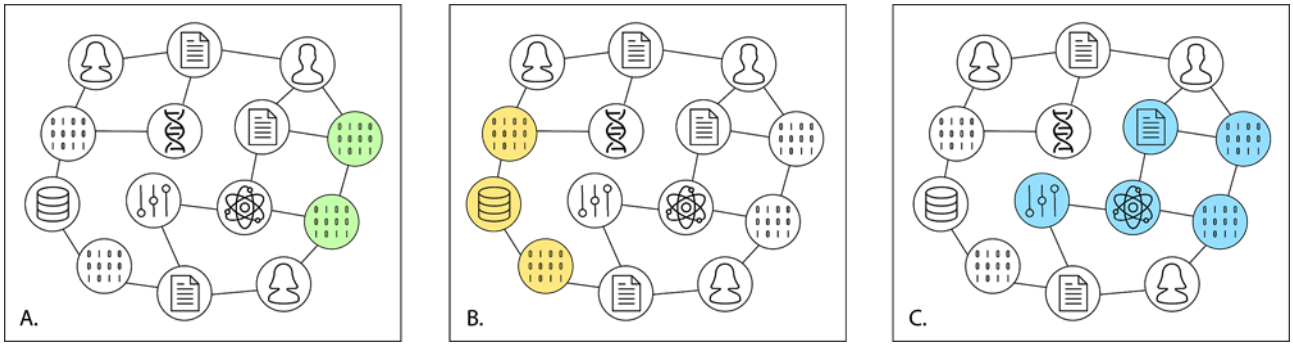

Figure 1 A schematic representation of the PID graph with digital objects connected by PIDs, showing three use cases: A: Different versions of software code, B: Datasets hosted by a particular repository, C: All digital objects connected to a research object.

Reuse across versions and parts (Figure 1 A)

Datasets and software (and to a lesser degree publications) are frequently versioned, and datasets can often be downloaded as subsets. Tracking the reuse (views, downloads, and citations) across versions and parts is a frequent use case and can lead to confusion in the community regarding proper versioning. An important publications user story is linking preprints and peer-reviewed publications.

Reuse of aggregated research outputs (Figure 1 B)

This might be the largest category of PID Graph user stories. We want to have a summary view of the reuse (via views, downloads, and citations) of all research outputs by a particular researcher, academic institution, data repository, or funder. This summary view can help demonstrate the impact of for example a researcher or repository.

Research objects (Figure 1 C)

Aggregate all scholarly resources that are linked together via a single publication, including underlying data and software used to generate the results, but also people, organizations, and funding involved in the work. Exploring the connections in and to a research object is currently very difficult, e.g. starting from a dataset

[1] https://github.com/datacite/freya/issues?q=is%3Aissue+is%3Aopen+label%3A%22PID+Graph%22

included in a research object, getting a list of publications that indirectly cite this dataset by citing the publication based on the data.

In January 2019, after launching the PID Forum[2], we migrated the user stories from GitHub issues to the PID Forum for broader feedback collection and discussion[3].

[2] https://www.project-freya.eu/en/blogs/blogs/work-through-your-pid-problems-on-the-pid-forum
[3] https://www.pidforum.org/c/pid-graph/17

3 Technical bases of the PID Graph

3.1 Initial work

After we had identified the most relevant user stories in our August 2018 workshop, we started to look into the technical architecture needed to support these user stories, starting with the existing infrastructure from the PID service providers among the FREYA partners. The existing REST APIs by PID providers Crossref, ORCID and DataCite describe the nodes in the PID Graph. The *identifiers.org* service by FREYA partner EMBL-EBI is less relevant, as it focuses on identifier resolution rather than returning metadata about resources with PIDs. Another important service that FREYA partners can use for the PID Graph is the Crossref/DataCite Event Data Service[4], which essentially stores the links (or edges) of a PID Graph, briefly describing the connected nodes with PIDs and optional metadata, but more importantly the edges, with provenance information such as the when and by whom the connection between two PIDs was made.

Most FREYA user stories ask questions that need two steps in a graph to answer, e.g. co-author graphs that need all co-authors of the research outputs of a particular researcher. To take advantage of the existing infrastructure, and as the user stories we were trying to address were not overly complex, we started work in September 2018 to adjust the existing PID provider backend infrastructures, including the Event Data services. These backend infrastructures used relational databases, Lucene-based search indexes (Solr or Elasticsearch) and REST APIs.

This exploratory work that we conducted between September 2018 and February 2019 showed us that we can use the existing PID infrastructure to address these user stories, but that we would be stretching the capabilities of our infrastructure, e.g. of relational databases and Lucene-based search engines. The user stories would typically need at least two queries, combining for example a query for PID metadata with an EventData query. We found that the query interfaces that we could expose to users would quickly become complicated, requiring a lot of bespoke development that would not scale well.

3.2 GraphQL

In March 2019 we thus started to explore a new technology that could help us implement PID Graph infrastructure: GraphQL. GraphQL is an open-source data query and manipulation language for APIs, and a runtime for fulfilling queries with existing data[5]. GraphQL, started in 2012 and open-sourced in 2015, has been widely adopted for building production webservices in terms of implementations, libraries, training and support. GraphQL provides a simple query interface that aligns well with the FREYA users stories, and can be used on top of existing PID infrastructure.

DataCite launched a pre-release version of a GraphQL API supporting the PID Graph in May 2019[6]. It became clear after a few months working with this GraphQL API that the technology is indeed a good fit for addressing PID Graph use cases. We have for example in April 2020 posted 15 example GraphQL queries directly addressing one of the FREYA PID Graph user stories identified in the August 2018 workshop[7].

The next step was then to turn this GraphQL API into production infrastructure available to FREYA partners, other EOSC projects and beyond. This work has now been completed, and on 6 May 2020 DataCite has released the first production version of its GraphQL API, with the following features:

[4] Dasler, R., & Cousijn, H. (2018, October 8). Are your data being used? Event Data has the answer! https://doi.org/10.5438/S6D3-K860
[5] https://graphql.org/
[6] Fenner, M. (2019, May 15). The DataCite GraphQL API is now open for (pre-release) business. https://doi.org/10.5438/QAB1-N315
[7] https://www.pidforum.org/c/pid-graph/17

1. Stable GraphQL schema, avoiding breaking changes going forward
2. TRL-8 infrastructure with redundant services, monitoring, documentation, and support, integrated into the rest of the DataCite infrastructure.
3. Significantly increased number of PIDs and associated metadata available for queries (see Figure 2)
4. Many changes in available query fields, addressing use cases raised since May 2019.

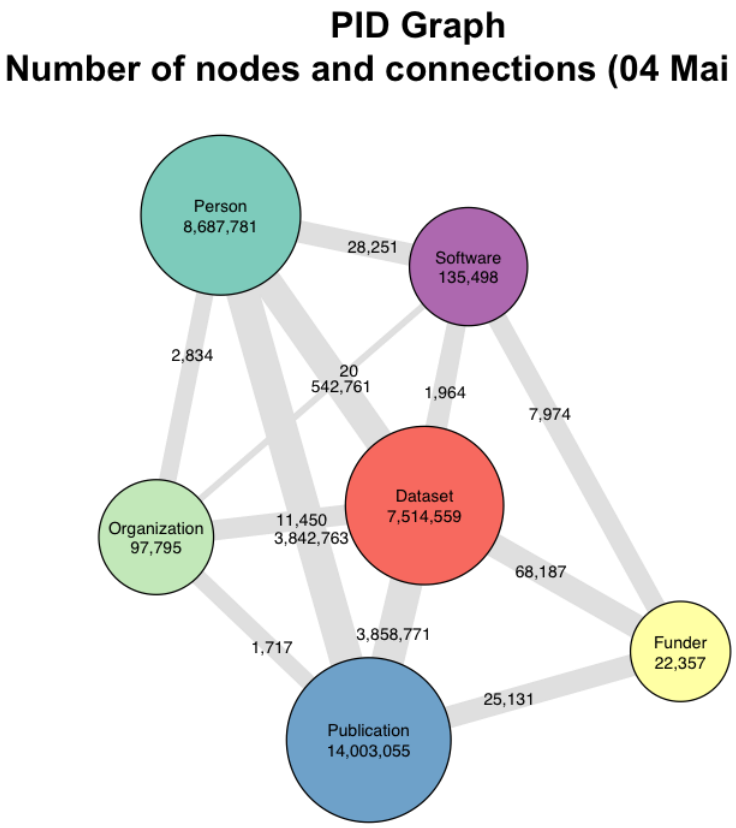

Figure 2 Nodes and connections in the PID Graph, May 2020

The nodes in the Graph shown in Figure 2 include all DOIs from DataCite, all ORCID IDs, all ROR IDs, all Crossref Funder IDs, and 8 ,577,869 DOIs from Crossref.

The service will be registered in the EOSC Portal Catalogue and Marketplace[8] after release.

[8] https://marketplace.eosc-portal.eu/

3.3 Federation

After launching the initial pre-release version of the DataCite GraphQL API in May 2019, in October 2019 we added federation based on the Apollo Federation technology that was announced in May 2019[9,10]. This federation technology allows us to integrate other GraphQL APIs or REST APIs with a GraphQL integration layer without storing any data at DataCite. We are doing this for ORCID, ROR, and Crossref Funder ID, and this will allow us to integrate other PID services going forward.

For Crossref DOIs we decided to do a tighter integration that better aligns with the deliverable *D2.4 Common DOI Search*. We are converting Crossref DOI metadata into the DataCite DOI metadata format and importing them into a common search index, enabling more sophisticated search results, e.g. relevance ranking across results from different DOI registration agencies. As of 27 April 2020, we have imported 8.5 million Crossref DOIs and included them in the GraphQL service, going forward we will import all remaining Crossref DOIs, making the Common DOI Search available not only via a web search interface (November 2020), but also via a GraphQL API (as well as a REST API).

[9] https://www.apollographql.com/blog/apollo-federation-f260cf525d21
[10] https://github.com/apollographql/apollo-server/tree/master/packages/apollo-gateway

4 PID Graph related outputs of FREYA

With the launch of the pre-release version of the DataCite GraphQL API in May 2019 FREYA is providing the core technical infrastructure for realizing the PID Graph. This work by FREYA WP2 is not associated with a specific deliverable, one important reason the PID Graph work is summarized in this document.

4.1 Jupyter Notebooks

We started to provide example Jupyter[11] notebooks of PID Graph user stories right after the launch of the DataCite GraphQL API. Jupyter notebooks easily integrate with the standardized GraphQL query interface, together providing an easy to use platform for addressing PID Graph user stories. Since the launch of the GraphQL API in May 2019 we have written dozens of notebooks, and have used them in webinars and workshops (see next section).

In May 2020 a contractor started subcontract work as foreseen in the FREYA Description of the Action to support the PID Graph work. He will be delivering 10 Jupyter notebooks addressing user stories from different disciplines and serving different stakeholder groups, and they will be fully documented and tested.

4.2 FREYA Work Package 2

FREYA work package 2 has two remaining deliverables in the six months of the FREYA project:

- D2.3 PID services registry (M31 June 2020)
- D2.4 DOI search service (M35 October 2020)

The PID services registry uses the GraphQL API to query for PID services registered with DataCite DOIs. The DOI search service for DOIs from multiple DOI registration agencies (DataCite, Crossref and others) is also built on top of the GraphQL API, supporting PID Graph functionality. This work will be reported elsewhere.

4.3 FREYA Work Package 4

FREYA WP4 has been developing the pilot applications showing uses of PID graphs in particular disciplinary areas. The core GraphQL infrastructure is now available to contribute to outputs in these deliverables in the remaining six months of the FREYA project:

- D4.5 Integration the PID Graph with the EOSC (M30 May 2020)
- D4.6 Emerging and new PID Graph resource types in disciplinary contexts (M36 November 2020)
- D4.7 Using the PID Graph: community workflows and discoverability services (M36 November 2020)

This work will be reported elsewhere.

[11] https://jupyter.org/

5 Training and outreach

Training and outreach materials have been created about the PID Graph since it was first launched in May 2019, and these have been focused on engaging early adopters through the use of Jupyter notebooks.

The DataCite GraphQL API production version with full accompanying documentation provides a timely opportunity to drive engagement and adoption of it. The publication of several user stories with GraphQL queries on the PID Forum[12] is a first step and will be followed up by additional Jupyter notebooks which are due to be written addressing user stories identified by FREYA partners.

Full documentation accompanying the GraphQL API has been made available via the DataCite support website which will support adopters of the technology[13]. This will be linked in the Developers section of the FREYA Knowledge Hub[14].

FREYA has been instrumental in launching the Research Data Alliance (RDA) Open Science Graphs for FAIR Data Interest Group[15] that first met in the April 2019 Philadelphia Plenary. This group is coordinating the various Open Science Graph activities and is co-chaired by Martin Fenner (DataCite, WP2 lead), Paolo Manghi (OpenAIRE) and others.

The PID Graph has been presented at several training and engagement events since it was initially launched in May 2019. These include:

- RDA UK Workshop, London, July 2019
- Open Science FAIR, Porto, September 2019
- Force11 2019, Edinburgh, October 2019
- RDA 14th plenary, Helsinki: FREYA co-located event “Connecting knowledge in the European Open Science Cloud”, October 2019
- PID NL workshop, The Hague, November 2019
- Software Graph Hackathon, London, December 2019
- IDCC, Dublin, February 2020

In addition to in-person events, the PID Graph has also been demonstrated in several online events including several FREYA organised webinars and events moved online due to the COVID-19 Pandemic. These include:

- FREYA Midterm Webinar, May 2019[16]
- DataCite UK Consortium Webinar, January 2020
- Research Data Canada Webinar, February 2020[17]
- Software Sustainability Institute Collaborations Workshop, March 2020

Forthcoming events include, EOSC Hub Week due to be held online in May 2020 and the FREYA end of project event in November 2020.

Additional webinars will be scheduled to communicate the PID Graph further, as further functionality is developed through to the end of the project, especially via the subcontract. These will be fully described in forthcoming deliverables D5.6 Final Training Materials and D5.7 Third report on the PID Forum.

[12] https://www.pidforum.org/c/pid-graph/17
[13] https://support.datacite.org/docs/datacite-graphql-api-guide
[14] https://www.pidforum.org/c/knowledge-hub/pids-for-developers/38
[15] https://www.rd-alliance.org/groups/open-science-graphs-fair-data-ig
[16] Recording available: https://www.youtube.com/watch?v=2FuPDzt7r8o
[17] Recording available https://www.youtube.com/watch?v=BALSvZIkDr8

6 Conclusions

FREYA is on track to implement a common technical basis for the PID Graph as proposed in the description of work, providing a unique and valuable resource for EOSC and beyond. Given the complex technical and social architecture needed to realize the PID Graph, the full potential will not be realized until after the end of the FREYA project, though the project's pilot applications will certainly show benefits in particular domains. The FREYA exploitation plan will elaborate on how PID Graph technology will be taken further and embedded in the offerings of the project partners, while the work on sustainability and the PID Commons will take account of the needs of the PID Graph in future.