Piloting the use of PIDGRAPH data in research.fi

Tommi Suominen[1], Laura Himanen[1], Lauri Hellsten[1] and Mike Bennett[2]
[1]CSC - IT Center for Science, Espoo, Finland
[2] DataCite, Hannover, Germany

Research.fi is a national service for collecting, integrating and disseminating information on research conducted in Finland. This includes information on publications and related research outputs (including in the arts), datasets, funding decisions and research infrastructures (Suominen & Rydman, 2021). There is a specific section for researcher profiles, focusing on researcher descriptions, their research activities, merits, education - moving towards an affiliation independent research CV of sorts (Nikkanen & Puuska, 2022). Research.fi also collects information on research news and open funding calls and hosts statistical information on the development of research resources and impact on a national level. A key aim is to be able to point to all of these research objects – by utilizing durable and unambiguous pointers – so called Persistent Identifiers (PIDs) (Meadows et al., 2019; Sompel et al., 2014).

Research.fi draws its information from mostly national sources, Cordis for EU funding information - but for the researcher profile a researcher can pull in data from Orcid (Haak et al., 2012) too. Research.fi's coverage of research objects themselves differs, based on the maturity of organizational data collection, and, for example. information concerning the publications of higher education institutions is very comprehensive as it is used as calculation criteria for HEIs basic funding, whereas information concerning research activities is incomprehensive as the motivation for collecting such information is quite low in research performing organisations. It is clear, however, that the coverage of information on the connection between these objects (Suominen & Rydman, 2021), such as publication – dataset references has serious shortcomings. This is primarily due to the legacy fact that these were earlier not prioritized in the data collection and consequently in the metadata of the different research objects. Already integrated data sources for connection information are described in Figure 1. However, a more complete inventory of connections maybe attained from Scientific Knowledge Graph (SKG) implementations such as the PIDGraph by DataCite (Cousijn et al., 2021; Fenner & Aryani, 2019). The PIDGraph is a collection of metadata about research objects, authors, publishers,

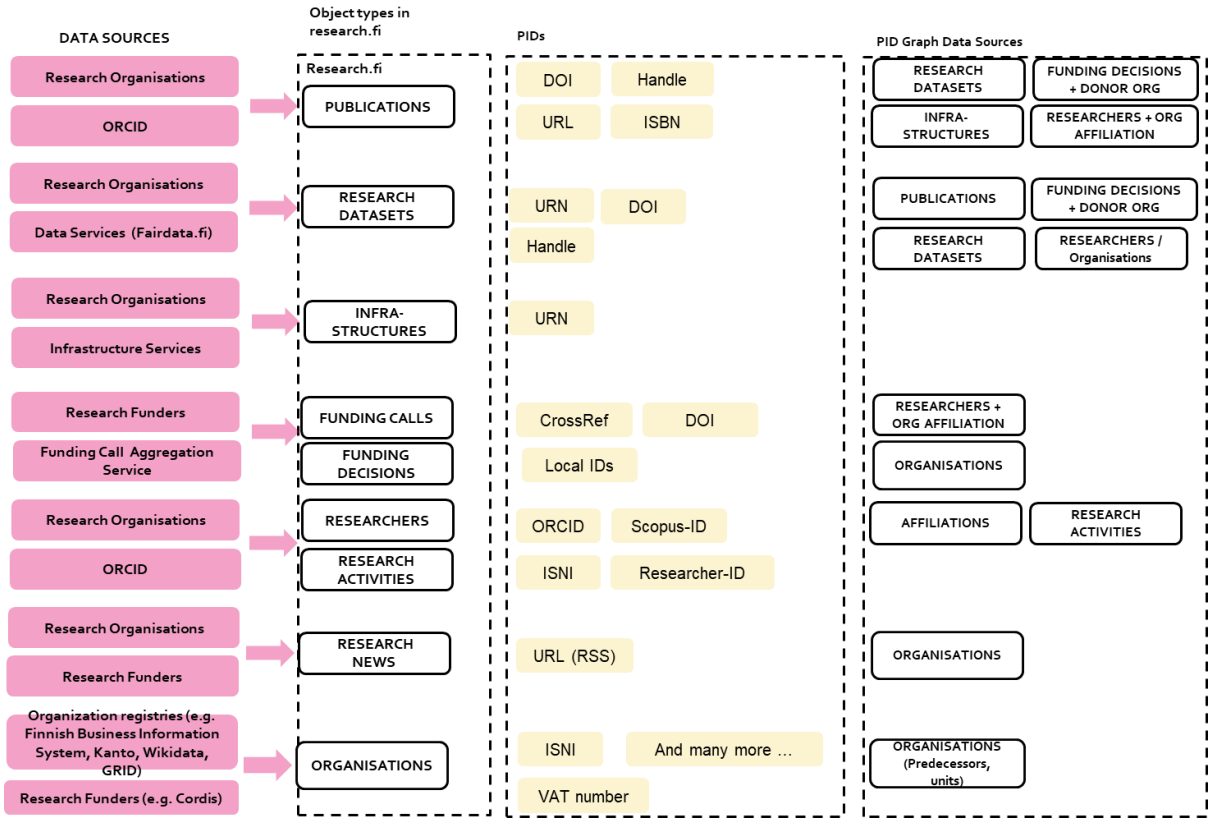

Figure 1 The different data sources for research.fi, by different research object types, which PIDS they may come with. Finally in the last column the types of connects that may come along with each object type (updated from Suominen & Rydman (2021)).

institutions, and other research related items, all identified by persistent identifiers, and a series of systems to expose that metadata and make it available for analysis and integration with other systems.

In order to harvest information on the connections between publications and datasets available in Research.fi, we have been looking into PIDGraph, first to find out the coverage of PIDGraph in regard to the content of Research.fi and secondly to see if it is possible to enrich research.fi with the information gained from PIDGraph. We aim at being able to link Research.fi publications to datasets and vice versa, publications to publications, and datasets to datasets, as well as enrich our current data with abstracts currently unavailable in research.fi. At the moment, out of the nearly 800 000 publications, only 21285 have an abstract in Research.fi

Initially, we used a PIDGraph data dump to retrieve information from PIDGraph, but the sheer amount of data, i.e. approximately 290 GB potentially amounting to millions of millions of rows in our database, proved difficult to work with. We decided to focus on namespaces for DOI, according to the folder names in the PIDGraph data dump. This resulted in 2218 rows, with only 53 of them such where both the DOI and also the related DOI were in our database. The next step of the pilot was to make a test search in the PIDGraph API by setting one DOI as a parameter. But as we have about 250 000 DOIs in Research.fi, and there is a limitation for 3000 requests per 5 minutes, it would have taken far too long. We then decided to test the GraphQL API with which we could precisely define the data we got as a result. The Research Organisation Registry (ROR) identifier (Lammey, 2020) was used as the parameter for identifying an organization. 51 out of the 71 organisations submitting publication information to Research.fi already have a ROR in our database. This means that 207164/277952 publications can be requested using a ROR. With regard to datasets, 36 of 42 organisations submitting datasets to Research.fi have a ROR in our database, but about half of the datasets in the research.fi database do not have any identifiers for organization. As a result, for 1331/5314 datasets we can use ROR as the parameter for the query and the remaining approx. 4000 datasets we need to query using the dataset DOI. The last test was to load all the data from the GraphQL API and the query parameter was ROR, so that still leaves about 20 organisations missing in this test. This resulted in 9424 relations between objects that we already have in our database, but for datasets we only found 12 relations for objects – this is much less than with the data dump. Also, we noticed that there were only 7 different relation types, compared to the data dump, where there were 21.

Development of the PIDGraph is ongoing, and planned enhancements include the development of country-level subgraphs, in part to address the use case of the Research.fi integration, as well as further enrichment of the relationship data between PIDs held in the graph, including ingestion of more sources of relationship data, and we anticipate increased matching of items inside Research.fi and metadata from the PIDGraph as these enhancements are developed. The testing of the PIDGraph is still a work in progress, and the results still need to be validated. We hope to achieve this by the time the euroCRIS conference takes place. The authors consider that sharing the exercise of ingesting research object relationship data is of high relevance to the CRIS community. We also aim to present the applicability of lessons learned for other potential adopters of the PIDGraph.

Acknowledgements

The work described in this paper has taken place in the context of the FAIRCORE4EOSC-project. FAIRCORE4EOSC has received funding from the EU's Horizon Europe research and innovation programme under Grant Agreement no. 101057264.

References

Cousijn, H., Braukmann, R., Fenner, M., Ferguson, C., van Horik, R., Lammey, R., Meadows, A., & Lambert, S. (2021). Connected Research: The Potential of the PID Graph. *Patterns*, *2*(1), 100180. https://doi.org/10.1016/j.patter.2020.100180

Fenner, M., & Aryani, A. (2019). *Introducing the PID Graph — FREYA*. https://doi.org/10.5438/jwvf-8a66

Haak, L. L., Fenner, M., Paglione, L., Pentz, E., & Ratner, H. (2012). ORCID: A system to uniquely identify researchers. *Learned Publishing*, *25*(4), 259–264. https://doi.org/10.1087/20120404

Lammey, R. (2020). Solutions for identification problems: a look at the Research Organization Registry. *Science Editing*, *7*(1), 65–69. https://doi.org/10.6087/KCSE.192

Meadows, A., Haak, L. L., & Brown, J. (2019). Persistent identifiers: The building blocks of the research information infrastructure. *Insights: The UKSG Journal*, *32*. https://doi.org/10.1629/UKSG.457

Nikkanen, J., & Puuska, H. M. (2022). Researchers' profiles in Finnish Research Information Hub. *Procedia Computer Science*, *211*(C), 206–210. https://doi.org/10.1016/J.PROCS.2022.10.193

Sompel, H. Van de, Sanderson, R., Shankar, H., & Klein, M. (2014). Persistent Identifiers for Scholarly Assets and the Web: The Need for an Unambiguous Mapping. *International Journal of Digital Curation*, *9*(1), 331–342. https://doi.org/10.2218/IJDC.V9I1.320

Suominen, T., & Rydman, W. (2021, January 27). *Research.fi is a national PID-graph implementation*. https://doi.org/10.5281/ZENODO.4494378