Grey Literature and Persistent Identifiers: GreyNet's Use Case

Dominic Farace, GreyNet International
Stefania Biagioni and **Carlo Carlesi**, GreyGuide ISTI-CNR, Italy
Chris Baars, DANS-KNAW, Netherlands

Abstract
The PID (Persistent Identifier) Project is a follow-up to the AccessGrey Project[1] carried out in 2019 in which an online survey was held among stakeholders in GreyNet's community of practice. Recipients were asked their opinions about persistent identifiers and grey literature. The focus now in this project is expanded to include the DOI for research outputs alongside the ORCiD for authors/researchers, and the ROR ID for research organizations. This project seeks to go beyond a straightforward compilation and linking of these PIDs by building the PID Graph and contribute to other PID-Graphs built by service providers like OpenAIRE[2]. In this case, the PID Graph seeks to demonstrate how persistent identifiers can further research in the field of grey literature; and, how they contribute in making research entities conform to the FAIR data principles: Findable, Accessible, Interoperable, and Reusable. PIDs and the PID Graph are also seen to serve in the digital transformation of grey literature and as such will contribute to education and training in this field of information. DataCite Commons[3] used in this project is a web search interface for the PID Graph. The results from queries directed to the PID Graph produced in this project will not only serve as a use case for GreyNet but will also provide a model for other communities of practice in grey literature.

Chapter Outline

1. Background – AccessGrey Project and Persistent Identifiers
2. Components and Data Workflow in GreyNet's PID Project
3. Implementation of the PID Graph
4. Persistent Identifiers and the PID Graph conform to FAIR data principles
5. Some Conclusions drawn from the PID Project

1. Background – AccessGrey Project and Persistent Identifiers[4]

In 2019, an online survey was carried out among GreyNet's community of practice in order to gain their opinions on the uses and applications of persistent identifiers for grey literature. Results from an online survey within the AccessGrey Project clearly indicate a positive opinion about persistent identifiers for grey literature. Emphasis in the AccessGrey Project focused on two of the four collections in the GreyGuide Repository – the GLP Collection of conference papers and the new RGL Collection of multiple grey literature document types. The search of records in the GLP collection enabled formulation of the questions used in the online survey pertaining to persistent identifiers in particular the DOI.

1.1 AccessGrey Questionnaire[5] - 'Persistent Identifiers and Grey Literature'

Q1. Persistent identifiers increase access to grey literature
Q2. Persistent identifiers serve as an incentive in the acquisition of grey literature
Q3. Persistent identifiers increase the citation of grey literature
Q4. Persistent identifiers allow for the preservation of grey literature
Q5. Persistent identifiers are vital in linking and cross-linking data
Q6. A DOI is a quality indicator that increases the value of grey literature

Q7. A repository or data archive that assigns DOIs to metadata records is more likely to attract content providers
Q8. Do you have an ORCiD or another author/researcher unique persistent identifier?
Q9. Does one or more of your publications have an assigned DOI?

Survey Population: 509		Survey Respondents: 56		Survey Results: 11%	
	Strongly Agree	**Agree**	**Uncertain**	**Disagree**	**Strongly Disagree**
Q1	31	19	6	0	0
Q2	17	22	15	2	0
Q3	31	19	5	6	1
Q4	23	23	8	2	0
Q5	33	15	7	1	0
Q6	17	19	13	6	1
Q7	19	26	7	3	1
	Yes	**No**	**Non-Applicable**		
Q8	37	13	6		
Q9	40	9	5		

Table 1: AccessGrey Survey Results

The results of the questionnaire, which constituted the first part of the project, were significantly positive regarding persistent identifiers and grey literature.

1.2. DOI an Incentive for Acquisitions

The results from the second part of the AccessGrey project however did not indicate that the minting of DOIs for research outputs would be a sufficient incentive for their acquisition in a repository, namely one that relies on self-archiving. During that project, new records entered in the RGL (Resources in Grey Literature) collection received a DOI and a system generated citation. However, fewer full-text metadata records were harvested during this part of the project than expected. This perhaps coincides with the response to the survey question (Q2) in which nearly 27% of the respondents were uncertain whether persistent identifiers serve as an incentive in the acquisition of grey literature.

2. Components and Data Workflow in GreyNet's PID Project

A persistent identifier (PID) is a permanent reference and unique label to an object that is independent of the storage location. The unique label ensures that the object can always be found, even if the name of the object or the repository changes. As a result, an object can always be found unambiguously on the basis of its PID. This is important for the long-term storage (archiving) of objects in a rapidly changing world.[6]

In short, PIDs
1. Provide the address of an object such as a landing page in a repository
2. Can be used to link objects and in so doing connect other associated metadata in a record
3. Unambiguously Identify objects even if they move to other systems and services
4. And are computer readable, demonstrating their interconnectedness with other research communities

The value of the PID (persistent identifier) not only provides a link to a digital object be it a person, publication, or organization, but also allows the metadata associated with the digital object to become connected. When that metadata itself is expressed as a PID, this further allows for the creation of a PID Graph that models FAIR data principles: Findable, Accessible, Interoperable, and Reusable. These four principles will be discussed further in the chapter, following the introduction of GreyNet's network service in which they are applied.

2.1 GreyNet International, Grey Literature Network Service[7]

In order to obtain optimal use of persistent identifies a sustained data infrastructure must be in place within a community of practice, one that facilitates a coherent data workflow. It is only then that a PID Graph can be constructed and implemented. In this section, we look at the various components in GreyNet's data infrastructure and then discuss how they are implemented within its workflow. It is Important to mention here that GreyNet's workflow as it applies to this project includes retrospective input.

2.2 Components of the Data Infrastructure integrated in the PID Project

■ **GreyGuide Repository[8] and Portal to Good Practices and Resources in Grey Literature**
GreyNet International collaborated with ISTI-CNR to construct the GreyGuide a web access repository, which would come to house its collections of accepted conference proposals (GLA), published conference papers (GLP) and author-researcher biographical records (BIO). To this end, open-source software was identified and incorporated, metadata templates were created to fit the three document types, and in 2017 these collections were fully online accessible.

■ **OpenDOAR, Directory of Open Access Repositories[9]**
The GreyGuide Repository is registered in the OpenDOAR Directory of Open Access Repositories[10]. It can be mentioned here that while the GLA and BIO collections in the GreyGuide Repository rely on self-archiving, records in the GLP collection are entered by the system manager.

■ **DOI, Digital object identifier[11] and DataCite.org[12]**
In 2018, GreyNet became a DOI minting service within DataCite and began assigning DOIs to its collection of Conference Papers in the GreyGuide Repository. Since it is this collection upon which our PID Project is based, the ORCiD and ROR IDs had to be included in the DOI records in order later construct the PID Graph and be part of other PID Graphs (e.g., DataCite Commons, OpenAIRE).

■ **ORCiD, Open Researcher and Contributor ID[13]**
Also, in that same year, ORCiDs were included in biographical records in the GreyGuide; and, an active campaign among GreyNet's author's and researchers was initiated – encouraging them to register an ORCiD if they did not yet have one. In order to facilitate this, a link to the ORCiD registry was provided[14].

■ **ROR, Research Organization Registry ID[15]**
In 2020 the ROR ID for research organizations was added as a metadata field in BIO records in the GreyGuide. By way of a search in the ROR Registry, ROR IDs of organizations could be online accessed and included in the records of those authors and researchers, whose conference papers are archived in the GLP collection as well as in their corresponding DOI records in DataCite. GreyNet has since applied for a ROR ID and awaits its assignment. A ROR ID unlike an ORCiD is not assigned separately but rather in interval batch-releases, once new records have been approved.
It is worthwhile to note that the ROR ID of an organization linked with a research output such as a conference paper or other grey literature document type might be perceived as

a quality indicator. If we look back to question (Q6) in the AccessGrey Survey, over 23% of the respondents were uncertain whether the DOI is a quality indicator that increases the value of grey literature. However, if DOIs were connected to their corresponding ORCiD and ROR IDs, there might be less uncertainty.

2.3 The Data Workflow implemented in the PID Project

Our PID Project team was formed bringing together human resources and expertise needed, namely the system management and development of the GreyGuide Repository; the communication and network management of the GreyNet Community, and the acquired knowledge and experience of the PID Graph. From early January 2021 through the first week of March 2021, GreyNet undertook three tasks integral to the PID Project.

First, to complete minting DOIs for its collection of conference papers in the GL-Series including those published in 2021. The collection now totals 443 conference papers with DOIs in DataCite that accounts for the population of our project. Other service providers, namely DANS EASY[16] for GreyNet's published datasets and the TIB AV Portal[17] for its conference video presentations also have assigned DOIs in DataCite; however, these are not included in the population of the project.

The **second** task that ran parallel with the first was the retrospective search and retrieval of ORCiD and ROR IDs that were added to both the DOI metadata records and their respective BIO records in the GreyGuide Repository. The retrospective task also included the input of biographical records on behalf of authors and researchers whose names appear in the GLP collection, but who had not yet submitted a BIO record. This was accomplished in part by retrieving biographical notes from previous conferences in the GL-Series preserved in GreyNet's inhouse archive and partly via Google searches.

A **third** ongoing task dealt with records that needed some modification in order to benefit the PID Project, such as
(1) An existing ORCiD in a record carries 16 digits but is not proceeded by [https://orcid.org/] and as such is not actionable; (2) An ORCiD is retrieved only to find the message 'No public information available', which makes it difficult if not impossible to confirm the identity of the author/researcher; and (3) When an author or researcher's organization is absent or unclear in a record, it becomes difficult or is not possible to assign a ROR ID using the ROR Registry.
While these and other such problems were few in number, the time required to correct them was disproportionate. Nevertheless, when a system and service rely on self-archiving and when a persistent identifier such as the ORCiD can only be acquired by the author-researcher – him or herself, then these tasks must be calculated in the workflow.

2.4 Compilation of Actionable Persistent Identifiers

Now that the complete collection of conference papers in the GL-Series has an assigned a DOI in DataCite, which incorporates their corresponding ORCiD and ROR IDs, the number of actionable persistent identifiers for our project is accounted for. And, this then allows for the construction of the PID Graph.

Conference Papers	Authors-Researchers	Research Organizations
GLP Collection: 443	BIO Collection: 238	BIO Collection: 238
443	146	180
100%	61.3%	75.6%

Table 2: Actionable PIDs compiled in the Project (as of *March 13, 2021*)

3. Implementation of the PID Graph

In an article published in early January 2021, GreyNet's attention was drawn to the benefits of connecting the various types of persistent identifiers in producing a PID Graph[18]. For our project, this includes the DOI, ORCiD, and ROR ID. It is expected that this PID infrastructure would further demonstrate the value of persistent identifiers and open the potential for more research - in our case, research in the field of grey literature.

To construct the PID Graph two elements are required:

(1) backend services that collect PID connections in a standardized way focusing on two PIDs that are connected. This is essentially building the elements of the graph;

(2) query interfaces that combine these connections with PID metadata. A technology that is highly suitable is GraphQL[19]. GraphQL is an open-source data query and manipulation language for APIs, and a runtime for fulfilling queries with existing data. This widely adopted query language provides a standardized interface that can be federated, making it easier to build client applications for the PID Graph. Applications built on top of the PID Graph allow users to explore the rich connections between PIDs and to address specific use cases. The PID Graph demonstrates that we can gain more from PIDs when we look at their connections – indicating that the sum is more than its parts.

3.1 Examples of the PID Graph

Below are examples of four PID Graphs drawn from GreyNet's store of persistent identifiers. Each graph is comprised of multiple resources (nodes) that are connected by lines (edges). In the first diagram, the PID Graph appears in horizontal format and depicts from a DOI perspective three publications connected with the authors and their respective organizations. In the second diagram, the PID Graph appears in cluster format and depicts from a DOI perspective the same three publications as in the first diagram; however, now they are connected with the authors' names and their respective organizations. In the third diagram, the PID Graph depicts from an ORCiD perspective an author and his respective organization linked to seven publications. One of the publications is further linked to three co-authors of whom only one organization is shown. And, in the fourth diagram, from a ROR ID perspective – a research organization is encircled and linked to a cluster of publications that is further encircled and linked to a number of authors. Three of the authors appear also linked to their own respective organization.

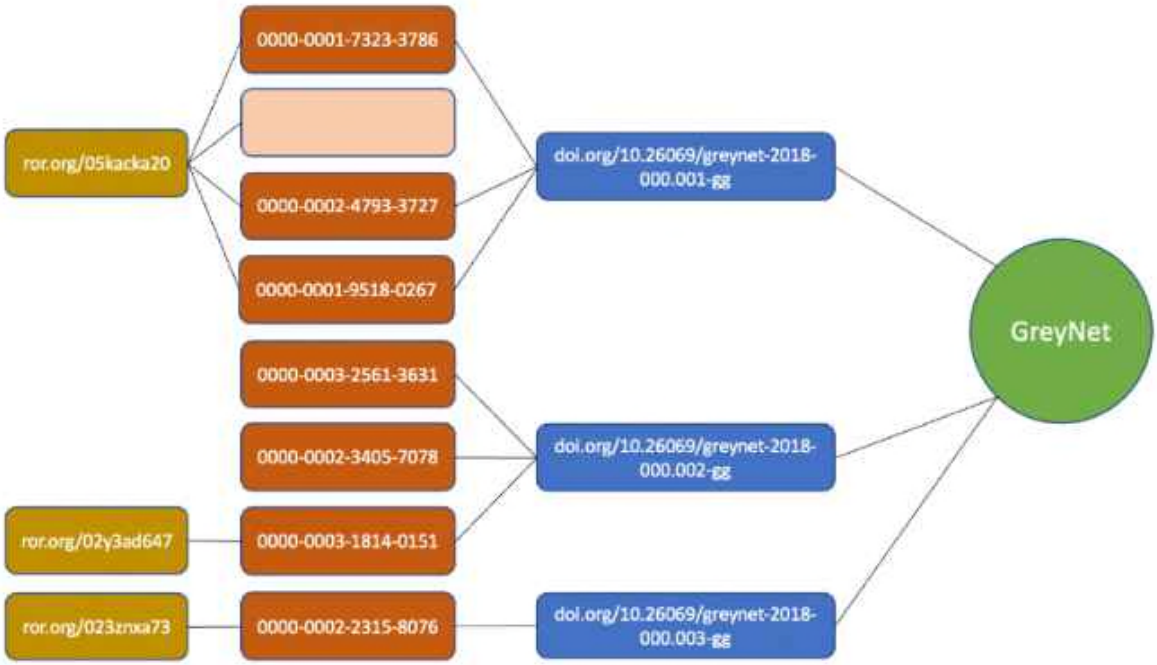

Diagram 1: PID Graph from a DOI perspective in horizontal format

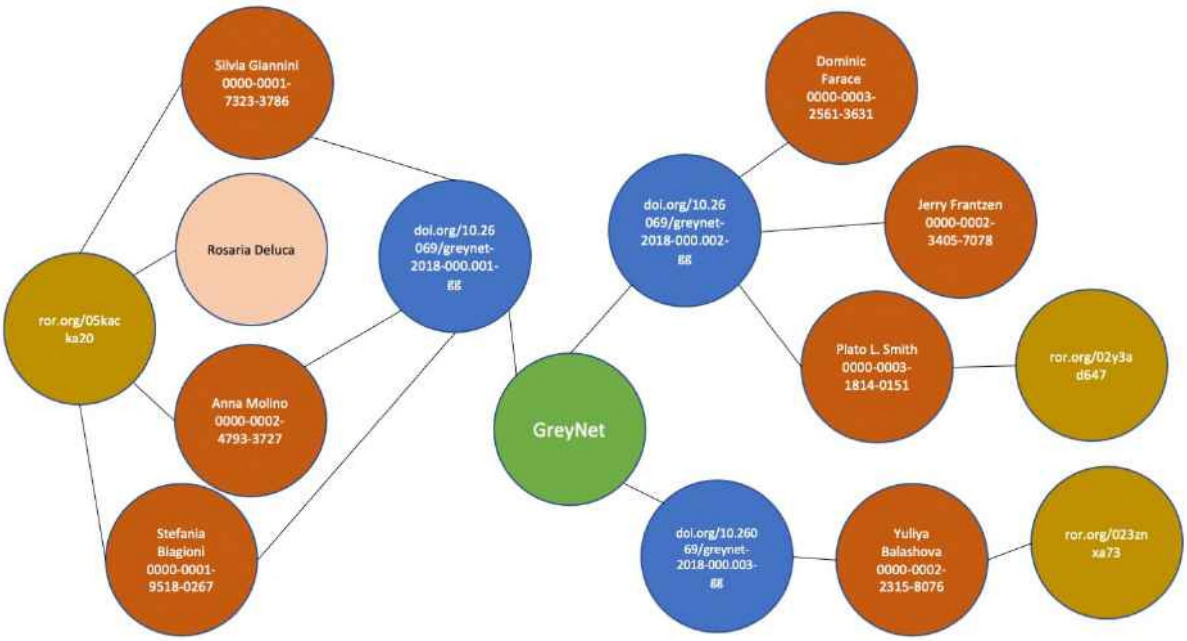

Diagram 2: PID Graph from a DOI perspective in cluster format

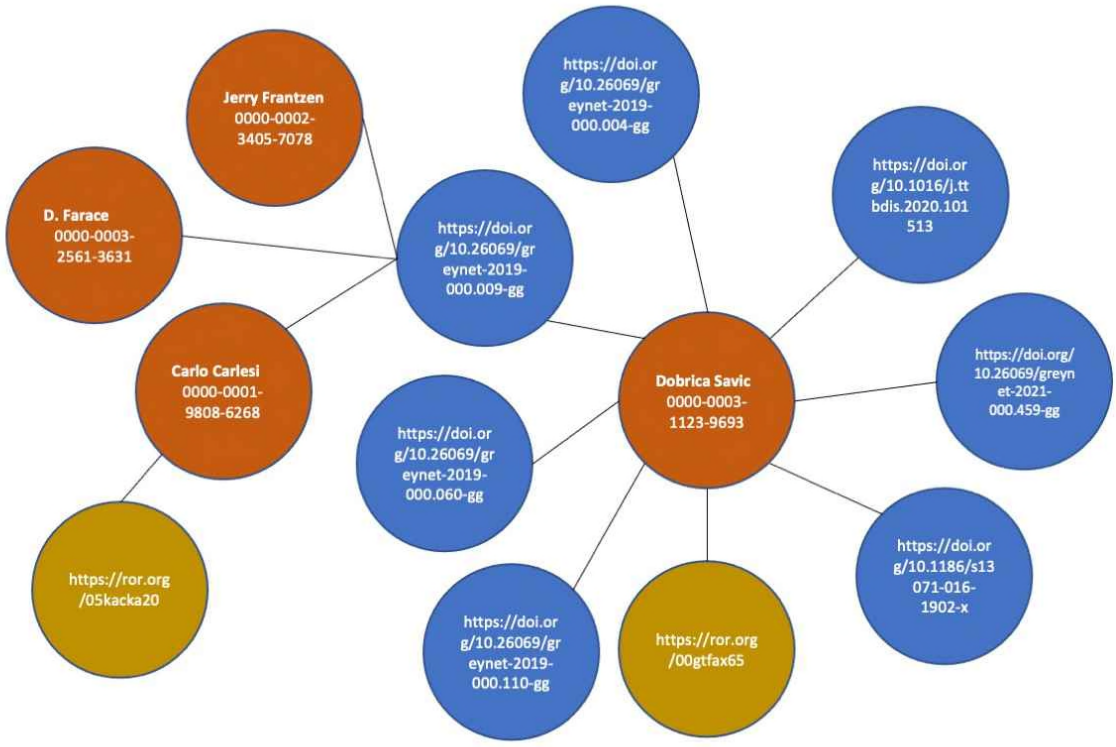

Diagram 3: PID Graph from an ORCiD perspective in cluster format

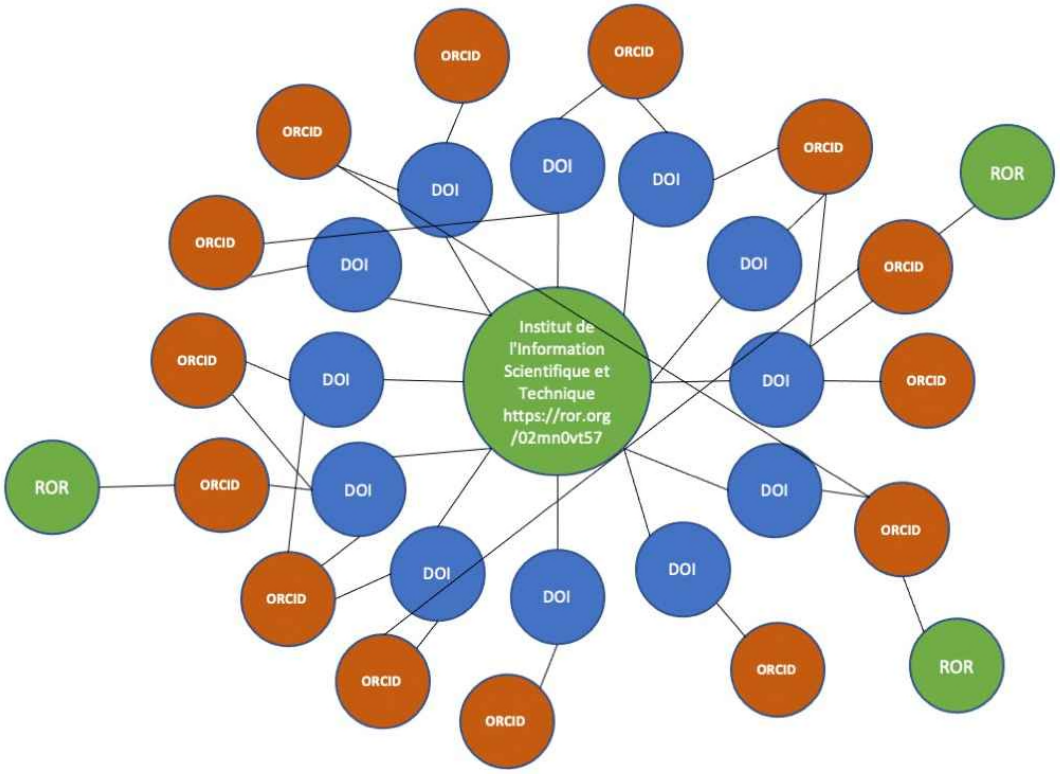

Diagram 4: PID Graph from a ROR ID perspective in cluster format

In the PID Graph, persistent identifiers are themselves the basic entities that are linked together; whatever they refer to is left implicit. This approach requires that the PID metadata are sufficiently rich to represent the relationships of interest and that the PIDs are of high enough quality. The advantage is that it becomes much easier to create graphs and to implement and scale rather than working with concepts and knowledge extraction.

4. PIDs and the PID Graph in relation to FAIR data principles

PIDs themselves allow for the guarantee of interconnected services from minting to linking onto access and preservation. When these services are situated in the workflow of a mature community of practice, they create a FAIR research environment.

[An extract abridged and revised from 'Connected Research: The Potential of the PID Graph'][20]

PIDs contribute in making research entities conform to the FAIR data principles; Findable, Accessible, Interoperable, and Reusable. By way of the PID Graph connections between different entities within the research landscape allow researchers to access new information. PIDs also play a role in the reusability of data by enabling rich metadata and their provenance to be associated with a digital object. PIDs provide the possibility to link entities long-term and enable information exchange by identifying persons and organizations over different services.

The overall PID infrastructure is made up of PID service providers, repositories, curation systems, aggregators, indexes, metadata, standards, and people. PIDs connect all of these elements, not only technically, via metadata and integrations, but also socially, via communities that have formed over decades or longer. The table below identifies the various types of PIDs and the maturity of their infrastructure. Since 2018, the ROR ID has moved from an emerging entity to a mature one.

Research Entity	PID Types Used	Maturity of PID Infrastructure
Publication	DOI, accession number, handle, URN, Scopus EID, Web of Science UID, PMID, PMC, arXiv identifier, BibCode, ISSN, ISBN, PURL	mature
Citation	OCI (secondary aggregation of information)	emerging
Conference	DOI, accession number	emerging
Researcher (or scholar)	ORCID IDs, ISNI (also DAIs, VIAFs, arXivIDs, OpenIDs, ResearcherIDs, ScopusIDs)	mature
Organization	DOI, ISNI, GRID, Ringgold IDs, ROR IDs	emerging
Data	DOI, accession number, handle, PURL, URN, ARK	mature

Table 3: PID Types and Infrastructure Maturity

GreyNet International now in its 28th year can be considered a mature research community socially. By including PIDs for objects, projects, persons, and organizations in the metadata, the technical maturity of GreyNet's infrastructure can now likewise be demonstrated. As a result of this ① sustainable connections can be made ② objects, projects, persons, and organizations become computer readable and understandable by other services like DataCite and OpenAIRE ③ A PID-Graph can be created and GreyNet information can also become part of other PID-graphs ④ Other services, like OpenAIRE PID-Graph and DataCite Commons[21] can be used to query or for purposes of analysis, and ⑤ It is a demonstration of FAIR-principles for grey literature. To include and expand on the FAIR principles, PIDs and metadata help ensure that the entities they refer to are

- usable and citable: pointing directly to an object, such as a specific item or a specific version of a dataset; hence increasing the usability of that object for researchers. It also helps them formally cite research outputs such as data and resources, which in turn facilitates reuse and helps increase recognition.
- Assessable: PIDs enable reliable measurement and prediction of impact, facilitating a more strategic approach to investment, driving maximum benefit, and ensuring that valuable resources are sustained.

5. Some Conclusions Drawn from the PID Project

Research in the field of grey literature will likely increase due to the incorporation and use of persistent identifiers. PIDs like other rich metadata can be counted and cross tabulated, enabling researchers to examine relationships in and among diverse types of data. As such, PIDs are actionable and can be used for new research. Furthermore, PIDs and the PID Graph can be seen not only to serve research in grey literature but also extend to new services in areas of education and training.

PIDs and the PID Graph are shown to have real value in defining GreyNet's position as a mature research organization by sustaining and leveraging its resources, by adhering to the FAIR data principles, and by signaling increased trust in grey literature beyond its own community of practice.

While the minting of a DOI was not of itself a sufficient selling point in the earlier AccessGrey Project for attracting content to a repository, the DOI now linked to the ORCiD and ROR IDs illustrated by the PID Graph may prove more effective. Also, while the AccessGrey Project laid the groundwork and direction for this PID Project, it is our understanding that implementation of the PID Graph will go even further to provide a new strategy and approach to research in the field of grey literature.

Linked References

[1] https://doi.org/10.17026/dans-zzf-cje3
[2] https://graph.openaire.eu/
[3] https://commons.datacite.org/
[4] https://doi.org/10.26069/greynet-2020-000.219-gg
[5] https://doi.org/10.17026/dans-zzf-cje3
[6] https://nl.m.wikipedia.org/wiki/Persistent_identifier
[7] http://www.greynet.org/
[8] http://greyguiderep.isti.cnr.it/
[9] https://v2.sherpa.ac.uk/opendoar/
[10] https://v2.sherpa.ac.uk/id/repository/9690
[11] https://en.wikipedia.org/wiki/Digital_object_identifier
[12] https://datacite.org/value.html
[13] https://en.wikipedia.org/wiki/ORCID
[14] https://orcid.org/register
[15] https://ror.org/
[16] https://easy.dans.knaw.nl/ui/home
[17] https://av.tib.eu/
[18] https://doi.org/10.1016/j.patter.2020.100180
[19] https://graphql.org
[20] https://www.cell.com/patterns/pdf/S2666-3899(20)30244-0.pdf
[21] https://commons.datacite.org/

Selected Resources

Introducing the PID Graph
https://www.project-freya.eu/en/blogs/blogs/the-pid-graph
The FREYA project
https://www.project-freya.eu/en
Connected Research: The Potential of the PID Graph
https://www.cell.com/patterns/pdf/S2666-3899(20)30244-0.pdf
The power of PIDs: Using persistent identifiers to link research outputs in the Netherlands
https://www.dpconline.org/blog/the-power-of-pids
NARCIS enriched with the first results of a PID-Graph
https://dans.knaw.nl/en/current/news/narcis-enriched-with-the-first-results-of-the-pid-graph
AccessGrey: Securing Open Access to Grey Literature for Science and Society
http://greyguiderep.isti.cnr.it/linkdoc.php?idcode=2020-GL21-007&authority=GreyGuide&collection=GLP&&langver=en
AccessGrey Online Questionnaire
file:///C:/Users/GreyNet/Downloads/Survey%20Results%20(Anonymous)%20(4).pdf
Data from "AccessGrey: Securing Open Access to Grey Literature for Science and Society"
http://greyguiderep.isti.cnr.it/linkdoc.php?idcode=2020-RGL01-002&authority=GreyGuide&collection=RGL&&langver=en