

Persistent Identifiers: Current Landscape and Future Trends

Research Data Canada
IDs Working Group, Standards and Interoperability Committee

Developed by the RDC stakeholder community, including:
[Mark Leggott](#), [John Aspler](#), [Reyna Jenkyns](#), [Brian Corrie](#), [Robyn Nicholson](#), [Jill Liu](#), [Carly Huitema](#)

10.5281/zenodo.7065515

Version 2.0, March 2022

Table of Contents

[Change Log](#)

[1. Introduction](#)

[2. The PID Framework](#)

[3. The PID Landscape](#)

[3.1 Researcher Identifiers](#)

[3.2 Research Outcome Object Identifiers](#)

[3.2.1 Publications](#)

[3.2.2 Datasets](#)

[3.2.3 Research Software](#)

[3.3 Organization Identifiers](#)

[3.4 Funding Identifier](#)

[3.5 Research Project Identifier](#)

[3.6 Equipment Identifiers](#)

[3.7 Physical Sample Identifiers](#)

[3.8 Other Identifiers](#)

[3.9 The Future of PIDs](#)

[4. Discussion](#)

[4.1 Governance and Sustainability](#)

[4.2 Recommendations](#)

[5. Acknowledgements](#)

Change Log

Version	Date	Description of Change
1.1	Aug 26, 2020	Cloned, no changes
1.2	Sep 28, 2020	Restructured, new sections/subsections introduced
1.3	Oct 16, 2020	Updated Sections
1.4	Nov 20, 2020	Updated and assigned all sections
1.5	Jan 11, 2021	Added PID Ecosystem diagram and text.
1.6	Feb 20, 2021	Additional text and sections added.
1.7	Feb 26, 2021	Final document team meeting version
1.8	Mar 3, 2021	Launch of public draft
1.9	Mar 9, 2022	Final draft version
2.0	Mar 25, 2022	Final Version

1. Introduction

The research enterprise generates a great deal of information, both in physical and digital form—from descriptive information about researchers, to publications, to datasets resulting from a research project. This information is scattered across many systems and technologies, including human resource systems, grant management systems, publication databases, data repositories, and web pages. In order to facilitate greater access to this wealth of information and data, and thereby increase the impact of researchers' endeavors, it is important to increase the FAIRness of these outputs: Findability, Accessibility, Interoperability, Reusability.¹ A foundational approach to increasing the FAIRness of research outputs is to uniquely identify them in a way that creates a sustainable network of relationships and access points.²

Persistent Identifiers (PIDs) are the anchors that facilitate links between the related pieces of information. Essentially, *IDs* are labels that refer to a specific entity in the information landscape, such as an object, organization, person, or dataset. For example, in the same way that a 'Person' record has the fields 'First Name' and 'Last Name,' it should be a best practice for each

¹ Mark D. Wilkinson et al. *The FAIR Guiding Principles for scientific data management and stewardship*. Scientific Data, 2016. <https://www.nature.com/articles/sdata201618>

² Christine Ferguson et al. *D3.1 Survey of Current PID Services Landscape - Revised*. Zenodo.org, May 31, 2018. <https://doi.org/10.5281/zenodo.3554255>

such record to uniquely identify the person, since there can be more than one author with the same first and last names. A *persistent ID* adds value by providing a *long-lasting reference to a digital object that gives information about that object regardless of what happens to it*.

PIDs serve two primary functions: first, PIDs uniquely identify an object, person or organization, facilitating unambiguous reference to that entity; and second, PIDs provide a mechanism to find entities over time, even if they change location. A PID system adds a third function: a framework (e.g. software and associated registry) for discovering objects described by a PID and doing something with them (e.g. viewing the object, getting usage statistics) in a sustainable way. The lack of PIDs, or the poor use/maintenance of them, stifles the power of discoverable or actionable information gained from linking multiple items, or objects, together - a fundamental characteristic of information systems. When done properly and according to agreed upon standards, the resulting network of resources (sometimes referred to as the *PID graph*) not only ensures access, but also increases the impact of specific research outputs.

PIDs are urgently needed in research to unambiguously locate, link, and cite research outputs (e.g. journal articles, data, and other research products such as samples, software, and formulas), as well as actors in the research process, such as authors, funders, and institutions. PIDs are now available for literature, data, samples, authors and more.

This paper, first published in 2016³ and updated in 2022, is intended to provide an overview of the current landscape of research PIDs, and will provide insights into their role in developing a more cohesive virtual research environment, supporting the preservation, discoverability, and reuse of research information. It also provides a series of recommendations—grouped by broad stakeholder group—designed to focus actions in the Canadian ecosystem to achieve concrete results quickly. The stakeholder groupings are: Research Funders; Universities and Research Centres; Science-based Government Departments; Repositories, Publishers and Infrastructure Providers; Researchers; and National Research Data Management organizations.

The aims of this paper are to:

1. provide readers with a better understanding of the role of persistent identifiers;
2. identify best practices in the international research data management community that facilitate the development of robust and sustainable systems for identifying and linking research outputs;
3. describe the current state of adoption of PIDs in Canada and internationally;
4. offer recommendations to the Canadian research community about the adoption of PIDs in various contexts.

³ Leggott, M., Shearer, K., Ridsdale, C., Barsky, E., & Baker, D. *Unique Identifiers: Current Landscape and Future Trends*. Zenodo.org, September 9, 2016. <https://doi.org/10.5281/zenodo.557106>

2. The PID Framework

The PID ecosystem comprises a wide range of actors and instruments that function to create standards and best practices, sustain software platforms, provide training and support, and facilitate policy development that reflects communities' interests in providing access to research outputs.

One can view the components of this ecosystem via three lenses:

1. The *Philosophy of PIDs* highlights the role PIDs play in the FAIR context, how they can be made more sustainable and actionable, and the role they can play in ensuring access to research data.
2. The *Policy of PIDs* reflects the desire for stakeholders to ensure that publicly-funded research outputs are FAIR, and may materialize in the form of funder mandates (e.g. requiring an ORCID for grant submissions), publisher workflows (e.g. assigning DOIs to accepted papers and datasets) or the adherence to recommended PIDs based on requirements defined by communities of practice.
3. The *Practice of PIDs* reflects the creation of standards for PIDs, whether in a general context (e.g. DOIs for publications), or a discipline-specific best practices (e.g. the use of RRIDs for cell lines used in a specific study), the development of platforms and tools to support the creation and maintenance of PIDs, and the training and facilitation efforts needed to make the use of PIDs an integral part of the research ecosystem.

3. The PID Landscape

There are many types and formats of identifiers used in the research and scholarly context. This report focuses on those IDs most relevant for the research data management landscape, and for which there is considerable agreement in the broader community. Other PIDs will be listed below, but will not be discussed in detail in this document.

As the PID landscape is large and complex, some basic concepts are defined here to assist the reader:

- The **ID** is a name that identifies (that is, labels the identity of) either a unique object or a unique class of objects, where the "object" or class may be an idea, physical [countable] object (or class thereof), or physical [noncountable] substance (or class thereof).⁴
- A **PID** is a long-lasting digital reference to an object that gives information about that object regardless of what happens to it.⁵
- A **PID organization** (or **service**) is an organization that provides a service (typically a software service) that when requested creates a unique PID (typically according to a standard) that can be used in a system.

⁴ Identifier. Wikipedia.org. <https://en.wikipedia.org/wiki/Identifier>

⁵ Persistent Identifier. CASRAI.org. <https://casrai.org/term/persistent-identifier/>

- A **URI** is a string of characters used to identify or name a resource on the Internet.⁶
- A **portal**, or **discovery system**, is an interface (typically a web page and associated software components) that provides access to a collection of IDs and associated URIs, generally via a search or browse option.

The ultimate goal of an effective and sustainable research data ecosystem is to provide the user (human or machine) with the ability to discover unique outputs from the research data lifecycle and link those outputs with the individuals and resources that were involved in their creation. It is through this network, or “graph” of linked data (i.e. the combination of metadata and associated data files), that all stakeholders in the research data community are able to answer the full range of questions they may have.

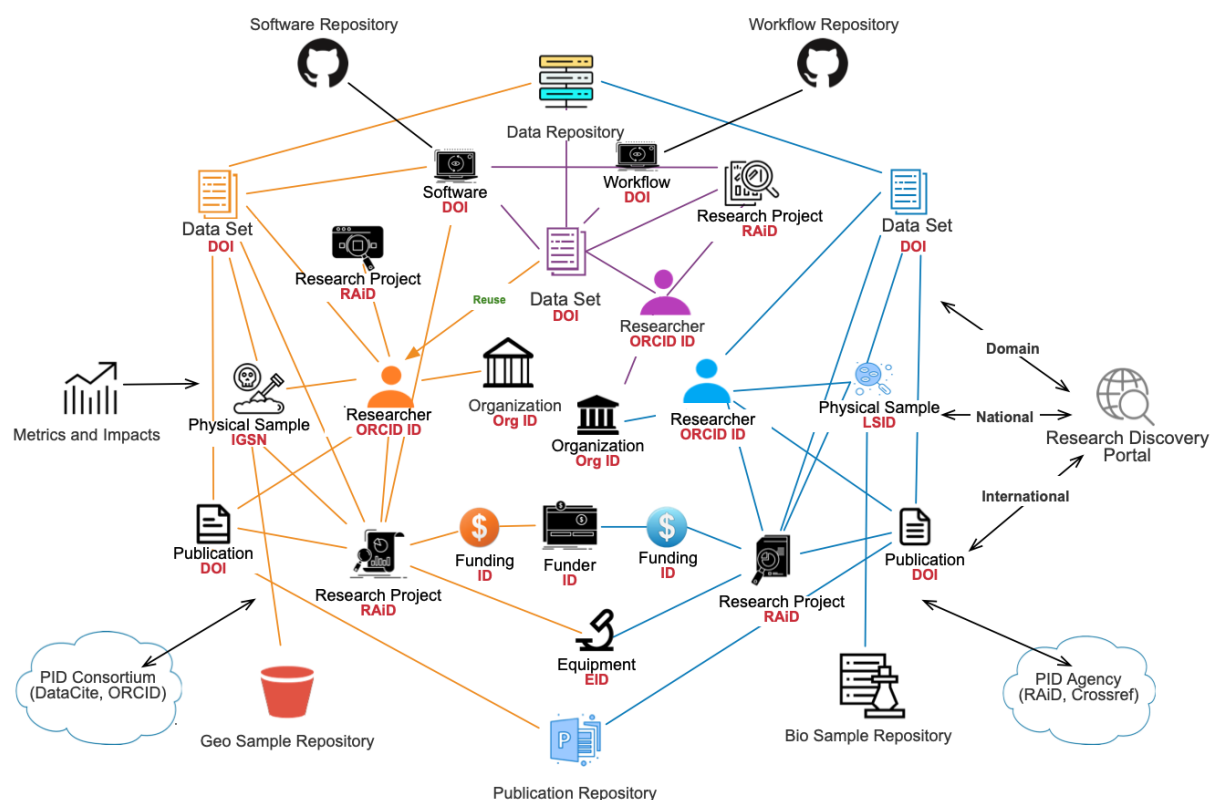


Fig.1 The PIDs Ecosystem

The PIDs Ecosystem diagram (Fig.1) is a simplified illustration of a sustainable and interoperable PID “ecosystem”, or PID graph, using three researchers (represented in orange, blue and purple). PIDs connect together the key elements involved in a research life cycle. As displayed in Fig.1, each element is assigned a unique ID facilitated by either a PID organization (e.g. DataCite, ORCID, RaiD, Crossef), or a non-affiliated PID service. In Fig.1, although the

⁶ Uniform Resource Identifier, CASRAI.org. <https://casrai.org/term/uniform-resource-identifier/>

research dataset, physical sample and research software that Researcher Orange used for their research and their research publication are preserved in different repositories, they are findable via one or more research discovery portals, made possible via the linked series of PIDs. For example, when Researcher Orange changes the institution they work for, or publishes their work under a different version of their name, their ORCID ID will stay the same and provide access to all that individual's outputs. Researcher Blue discovers Researcher Orange's research project, by using the information in its RAiD record, linked via the same equipment they used in their efforts. Researcher Blue is able to find Researcher Orange without any ambiguity with other researchers who have the same name because Research Project Orange is connected to Researcher Orange's ORCID ID. Researcher Orange is able to find a dataset produced by Researcher Purple via a search in an appropriate discovery portal. That dataset is also linked to the research software that both Researcher Orange and Researcher Purple use to analyze and produce their final datasets.

Without access to a rich discovery portal and open access to the research outputs, researchers are compromised in the ability to undertake effective research that builds on the efforts of others. The ability to derive metrics from the metadata in this linked ecosystem of PIDs also provides researchers, as well as other actors in the ecosystem, with the ability to provide attribution for all of a researchers' outputs, thereby ensuring a higher impact for that research.

Digital Object Identifiers (DOI)⁷ are the most commonly used type of PID in the scholarly communication context, and are a special case given their use in so many aspects of the research ecosystem. This section will describe the special case of DOIs, their development, and how they are used. More information about specific uses of DOIs is covered in subsequent sections.

The DOI system is managed by the International DOI Foundation (IDF)⁸, which provides oversight to DOI registration agencies and maintains the DOI resolver. DOIs must be assigned through registration agencies, such as Crossref and DataCite, which provide the infrastructure for organizations to manage DOIs. According to the IDF, over 230 million DOI names have been assigned to date.⁹ DOIs can be assigned for publications, datasets, and other research outputs, such as software, images, samples, and audio/video resources. Journal articles represent the majority of DOIs, with datasets rapidly becoming the second most common use of DOIs. Although most commonly assigned to research outputs, DOIs can be assigned to any type of object on digital networks, even if that object represents a real-world entity (such as a physical instrument or a funding organization).

A DOI contains two components: a *prefix* and a *suffix*, separated by a slash.

⁷ DOI Handbook: Introduction. DOI.org. https://www.doi.org/doi_handbook/1_Introduction.html

⁸ The DOI System. DOI.org. <http://www.doi.org/index.html>

⁹ Factsheet: Key facts on Digital Object Identifier Systems. DOI.org
<http://www.doi.org/factsheets/DOIKeyFacts.html> Accessed October 7, 2020.

- A DOI prefix designates a namespace for a registrant, such as an institution or a repository. All prefixes start with “10.” and include only numbers.¹⁰
- A DOI suffix is a string containing letters, numbers, and other characters.¹¹

For example, in the DOI 10.1038/nature19057, 10.1038 is the prefix and nature19057 is the suffix. Because together they constitute the DOI, the combination of the prefix and suffix must be unique within the DOI system.

A DOI is usually presented as a URL, for example <http://doi.org/10.1038/nature19057>. The format starting with “doi:”—for example, doi:10.1038/nature19057—is also commonplace; however, since this does not always resolve as a link in web browsers, DOI registration agencies now recommend that DOIs be formatted as URLs for usability.¹²

The DOI system implements and builds upon the Handle System, which provides persistent identifiers and resolution for digital objects.¹³ The DOI system adds additional components to the Handle System to provide a more seamless and functional approach to the provision and resolution of PIDs, hence the close association between the two. Specifically, the DOI system adds metadata to the Handle system, with specific metadata schemas being managed by registration agencies.

3.1 Researcher Identifiers

3.1.1 Background

Researcher Identifiers, also known as Author Identifiers or Scholar Identifiers, are alphanumeric strings that establish a unique identity for a given author or creator. Researcher IDs (RIDs) are becoming more critical in the scholarly communication ecosystem. As the number of scholars grows, and there is an increasing likelihood of identical names, the resulting confusion about which author one is referring to will also grow. Additionally, a researcher may publish using several variations of their name over time, or they may change their name or institutional affiliation, leading to their work being incorrectly associated with the wrong author. RIDs help to disambiguate researchers by connecting a researcher with a unique identifier that is associated with them throughout their entire career. These PIDs also greatly facilitate the tracking or linking of scholars and their outputs, whether by people or by software agents and systems.

In the last few years, ORCID Identifiers (ORCID iDs) have emerged as the gold standard open, international, and community-driven RID, as discussed below in section 3.1.4 (Implementation

¹⁰ *DOI Basics, DataCite Support*. DOI.org. <https://support.datacite.org/docs/doi-basics> Accessed December 8, 2020

¹¹ *DOI Basics, DataCite Support*. DOI.org. <https://support.datacite.org/docs/doi-basics> Accessed December 8, 2020

¹² *DataCite DOI Display Guidelines*, DataCite.org, accessed December 8, 2020.

<https://support.datacite.org/docs/datacite-doi-display-guidelines>

¹³ *DOI System and the Handle System*. DOI.org. <https://www.doi.org/factsheets/DOIHandle.html> Accessed December 8, 2020

and Impact). Most other RIDs are proprietary or tied to a specific private vendor system or organization (e.g., Scopus IDs from Elsevier, Researcher ID from Clarivate, Google Scholar ID).

3.1.2 Technical Description

Over the years, different actors or systems have utilized different types of RIDs, including publishers, archives, libraries, commercial aggregators, database providers and funders. Some examples of these are:

- Scopus Author ID - used in the Elsevier Scopus database (e.g. - <https://www.scopus.com/authid/detail.uri?authorId=16468267100>)
- Google Scholar ID - used in the Google Scholar database (e.g. - <https://scholar.google.com/citations?user=H-M7ztAAAAAJ>)
- VIAF Number - The Virtual International Authority File is an aggregation of national authority data and provides a single unique ID for specific individuals. (e.g. - <https://viaf.org/viaf/104750011/>)

The problem with most of these IDs is that they are system-specific and often maintained by a single commercial organization for a specific internal purpose; therefore, they do not have the potential to be broadly adopted and cannot evolve into truly comprehensive services. In some cases, commercially maintained IDs (e.g., Ringgold IDs for institutions) have historically been an important part of otherwise open ID systems (i.e., ORCID has used Ringgold, among other organization identifiers, in the ORCID Registry), an issue which needs attention to ensure a sustainable and fully open ID ecosystem. For RIDs to be truly effective and powerful, they must be broadly and publicly accessible, and be both human and machine-readable.

ORCID has both a Public and a Member API.¹⁴ In the former case, anyone with an ORCID iD can build a public tool and use it to interact with users and their public information in the ORCID Registry. Examples include the ability to ask a user to authenticate their identity with ORCID, to collect a list of their publications, or to display an authenticated ORCID iD in a web profile. In contrast, ORCID member institutions can create institutional tools that, with a user's authorization, allows that institution to become a "Trusted Organization."¹⁵ This enables Member API functionality, including the ability to read certain hidden information, adding and updating information in an ORCID Record, and automatic synchronization via webhooks. Examples include asserting and updating a user's employment information or publication metadata on their behalf, or automatically collecting new publications as they are added to an ORCID Record.

¹⁴ *Integration and API FAQ*. ORCID.org. <https://info.orcid.org/documentation/integration-and-api-faq/> Accessed March 10, 2022.

¹⁵ *Trusted organizations*. ORCID.org. <https://support.orcid.org/hc/en-us/articles/360006973893-Trusted-organizations> Accessed March 10, 2022

3.1.3 Use Cases

Some typical use cases for RIDs are:

1. A publisher or repository would like to validate the name of an author and distinguish them from other authors with similar names.
2. A researcher would like to retrieve a list of publications by a colleague in a specific subject area. The colleague has changed institutions several times in her career.
3. A university administrator would like to assess the intellectual impact and output of a department or research group to determine their areas of strength.
4. A funder would like to retrieve all the publications by the researchers associated with a specific grant.
5. A research hospital, institute, or network would like to know more about their affiliated scholars' university affiliations.

3.1.4 Implementation and Impact

In the last decade, ORCID has developed a global solution for RIDs—ORCID iDs—which have gained traction internationally and are being integrated into institutions and systems across the world. At the time of the 2016 version of this paper, there were just over 2,250,000 iDs assigned to researchers globally. At the start of 2022, there are just over 14,000,000 ORCID iDs.¹⁶ Of these, there are at least 160,000 ORCID iDs associated with scholars at Canadian institutions. The ORCID Organization (properly referred to as *ORCID*) provides researchers with a unique alphanumeric ORCID identifier string (the *ORCID iD*), a space to track their scholarly activities (the *ORCID Record*), and the infrastructure to store and exchange information associated with the full collection of ORCID iDs and ORCID Records (the *ORCID Registry*).¹⁷

ORCID iDs are the de facto standard for RIDs. Individuals can register for an ORCID iD and populate their ORCID Record for free; however, the infrastructure of ORCID is sustained through institutional membership, often in local consortia. In Canada, this consortium is called “ORCID-CA” and is administered by the Canadian Research Knowledge Network (CRKN).¹⁸

The true value of ORCID lies in the space between individuals and institutions - while ORCID enables individual scholars to disambiguate themselves from others with the same or similar names, institutional members can integrate ORCID into their systems using the ORCID API (e.g., research management, digital repositories, grants and funding systems, digital publishing), which enables a more robust transfer of information between people, places, and systems.

ORCID's focus is on connecting people and research across disciplines, borders, and time. They are using ORCID iDs to more easily connect these elements of research, which have traditionally been a challenge for the community. ORCID connects person identifiers with key

¹⁶ ORCID maintains an active *ORCID Statistics Page*, which is updated regularly. ORCID.org. <https://orcid.org/statistics>. As of May 24, 2022, there are 14,109,428 ORCID iDs globally.

¹⁷ *What is ORCID?* ORCID.org. <https://info.orcid.org/what-is-orcid/>

¹⁸ *ORCID-CA: Home*. CRKN-RCR.ca. <https://www.crkn-rcdr.ca/en/orcid-ca-home>

workflows, from manuscripts to grant applications, encouraging the research community to interact with features such as tools, use cases, documentation, examples, and open-source code.

ORCID allows researchers to control the information stored in their own ORCID Record¹⁹ about themselves, their institutional affiliations, research funding, research outputs, and more, as well as to determine which information is public versus private.²⁰ This last feature is important because users can search the publicly available part of the registry for first names, last names, ORCID iDs, institutional affiliations (text string or organizational ID), and keywords. There is no 'browse' capability, but if the user has a clear idea of what or whom they are looking for, the platform is robust - and ORCID also provides a more complex URL-based search tool to help download certain search results.²¹ The addition of institutional login using Shibboleth also provides an additional level of "institutional authority" to the maintenance of ORCID iDs.

Examples of resolvable ORCID iDs, with varying levels of complexity in each ORCID Record:

- A fictional ORCID Record:  <https://orcid.org/0000-0001-5727-2427>
-  <https://orcid.org/0000-0001-6133-4045>
-  <https://orcid.org/0000-0001-7055-4357>
-  <https://orcid.org/0000-0003-1392-7799>
-  <https://orcid.org/0000-0003-3888-6495>

In the last year, ORCID has also released a Member Portal,²² which contains detailed metrics²³ and a members-only tool enabling institutions to easily add affiliations with scholar permission.²⁴

As discussed, as a global, open, community-driven not-for-profit, ORCID remains sustainable (and ORCID iDs remain persistent) through the support of institutional membership. Individual scholars using ORCID do so for free, but universities, government departments, publishers, funders, and more pay to access the infrastructure and software tools to integrate ORCID into their local systems—with more affordable memberships afforded to public and non-profit institutions. In exchange, members gain access to specific Premium Member API features (i.e., the ability to become a trusted organization and to read certain private information or add metadata to an ORCID Record with a user's permission, as well as webhooks) and technical support, as well as other member benefits. In addition, a growing number of funders require the adoption of ORCID iDs by their researchers or have explicit ORCID policies.²⁵ As of January

¹⁹ *Privacy Policy*. ORCID.org. <https://info.orcid.org/privacy-policy/>

²⁰ *ORCID Trust*. ORCID.org. <https://info.orcid.org/orcid-trust/>

²¹ Paula Demain. *How do I find ORCID record holders at my institution?* ORCID.org, February 17 2020. <https://info.orcid.org/faq/how-do-i-find-orcid-record-holders-at-my-institution/>

²² *ORCID Member Portal*. ORCID.org <https://info.orcid.org/documentation/member-portal/>

²³ *Member Reporting*. ORCID.org <https://info.orcid.org/documentation/features/member-reporting/>

²⁴ *Affiliation Manager*. ORCID.org <https://info.orcid.org/affiliation-manager/>

²⁵ *Funders' ORCID policies*. ORCID.org, accessed March 10, 2022. <https://info.orcid.org/funders-orcid-policies/>

2021, ORCID has been integrated into over 900 systems, with more in the works.²⁶ Publishers also increasingly require ORCID iDs to represent the unique ID of an author in their systems.²⁷ There are currently about 1,170 members in ORCID across 51 countries and 23 national or regional consortia. Notably, in 2020, The Public Knowledge Project (PKP), ORCID, ORCID-CA, SciELO, and the ORCID US Community collaborated to improve and promote the ORCID plugin for Open Journal Systems (OJS), which is now in use or in preparation at over a dozen ORCID-CA member institutions.

ORCID-CA, which has 41 members as of April 2022, was formally launched in 2017. In May 2016, two workshops were organized in Ottawa and Toronto to raise awareness of ORCID iDs and to recruit institutional members. The aim was to discuss the importance of RIDs in enabling a strong research infrastructure and to demonstrate the value of a cross-sector approach and application of ORCID iDs in research infrastructure with various communities. Many stakeholders participated including CASRAI, CARL, CRKN, NRC, RDC, the Tri-Agency, and others,²⁸ leading to the publication of a Joint Statement of Principles in 2017.²⁹

Since this time, ORCID-CA has grown substantially.³⁰ With 41 members, 46 member integrations, and over 160,000 ORCID iDs associated with scholars at Canadian institutions, ORCID-CA supports members bilingually in the promotion, adoption, and integration of ORCID across Canada, as well as through a Community of Practice and digital resources for members. These members are often University Research Libraries, but can include government departments and crown corporations, other parts of a university such as a Research Office, and broader research networks and hospitals. The members of the original stakeholder group that published the Joint Statement of Principles went on to form two governing bodies: a Governing Committee (made up of and elected by ORCID-CA members) and an Advisory Committee (made up of key non-member stakeholders). Recently, the latter committee expanded its scope to include the work of the DataCite Canada Consortium and to provide advice about PIDs more broadly across Canada, with the aim of broader national PID implementation. It is now called the “Canadian PIDs Advisory Committee” (CPIDAC) and includes representation from the Tri-Agencies, CARL, Scholars Portal, The Digital Research Alliance of Canada (The Alliance), PKP, CRKN, Compute Canada, CANARIE, and more.

²⁶ Meadows, Alice, *Collect & Connect: Turning ORCID's Vision into Reality*. ORCID.org, May 31, 2016.

<https://info.orcid.org/collect-connect-turning-orcids-vision-into-reality/>

²⁷ *Manage My Author Profile*, Elsevier.com, accessed March 10, 2022.

<https://www.elsevier.com/solutions/scopus/support/authorprofile>; as well as *ORCID in Publications*, ORCID.org.

<https://info.orcid.org/requiring-orcid-in-publications/>

²⁸ Claire Appavoo et al. *ABC Project 2 - Launching an ORCID Consortia in Canada*. November 15, 2016.

<https://www.slideshare.net/CASRAI/abc-project-2-launching-an-orcid-consortia-in-canada-clare-appavoo-geoffrey-harder-mark-leggott-68973528>

²⁹ *Connecting researchers and research - Joint Statement of Principle and Proposal for ORCID-CA Consortium*. 2016.

<https://www.crkn-rcdr.ca/sites/crkn/files/2021-03/ORCID%20CA%20Joint%20Statement%20of%20Principles-Final%20%28EN%29.pdf>

³⁰ For a list of members as of December 2020, please see *Persistent Identifiers in Canada: ORCID-CA and DataCite Canada - White Paper*. <https://alliancecan.ca/white-papers/persistent-identifiers-in-canada-position-paper>

Membership in ORCID-CA is divided into two separate fee categories: The ORCID License Fee is currently set at \$3,600 USD annually per member for a consortium of 35-60 members. The ORCID-CA membership fee, supporting the local work of ORCID-CA, is roughly \$120,000 CAD for 2021-2022, and is divided equally among members (~\$3,200 CAD annually based on a total of 38 members in May 2021). Consequently, annual ORCID-CA membership fees can range from roughly \$7,500-\$10,000 CAD per member depending on factors such as the USD to CAD exchange rate, the total ORCID-CA membership, and applicable taxes. To facilitate adoption in Canada, the Digital Research Alliance of Canada provides funding to CRKN to reduce the membership fees for both ORCID-CA and DataCite Canada. Another cost to consider includes developer time and funds to build and support ORCID integrations. Many large vendor products from major research organizations have already integrated ORCID into their systems, meaning that members could move forward without development costs; however, those vendor systems and tools may themselves be financially out of reach to many organizations.

Beyond Canada, there are 22 other local ORCID consortia, including in France, Australia, Germany, Japan, South Africa, the UK, Brazil, the US, and more.³¹

In Germany, for example, the ORCID DE project was launched in February 2016 to support German universities and research institutions that are considering implementing ORCID in a coordinated and sustainable approach. At the time of launch, the project aimed to address organizational, technical, and legal issues in addition to providing a central contact point for universities and research institutions. ORCID DE also focuses on the cross-linkage and use of ORCID iDs in open access repositories and publication services. ORCID DE's project launch partners are the Helmholtz Open Science Coordination Office at the GFZ German Research Centre for Geosciences, the German National Library (DNB) and the Bielefeld University Library. The project was initiated by the German Initiative for Network Information (DINI).³² Today, there are 64 institutional members in the ORCID DE consortium.³³

Similarly, Australia's ORCID consortium has 42 members (38 universities, plus CSIRO, the Australian Nuclear Science and Technology Organization, the Australian Research Council, and the National Health and Medical Research Council).³⁴ In some countries, such as Italy, almost every researcher has an ORCID iD, facilitated by the efforts of national funders and other research agencies working together with ORCID.³⁵ In New Zealand, the Ministry of Business, Innovation and Employment covers the costs of institutional membership in the New Zealand ORCID Consortium (administered by the Royal Society Te Apārangi), which currently boasts 51

³¹ *ORCID Consortia Members*. ORCID.org. <https://orcid.org/consortia>

³² *Announcing the ORCID DE project to foster ORCID adoption in Germany*. ORCID.org. <https://info.orcid.org/announcing-the-orcid-de-project-to-foster-orcid-adoption-in-germany/>.

³³ ORCID DE. ORCID.de. <https://www.orcid-de.org/>

³⁴ *ORCID Members*. AAF.edu.au. <https://aaf.edu.au/orcid/members/>

³⁵ Mennielli, Michele, Andrea Bollini, Josh Brown, and Susanna Mornati. *Identify to simplify: improving interoperability at a national level with the ORCID HUB*. EUNIS, 2016. http://www.eunis.org/eunis2016/wp-content/uploads/sites/8/2016/02/EUNIS2016_paper_17.pdf

members³⁶, for up to 99 potential members.³⁷ In addition, the NZ Consortium has developed ORCID-compatible software tools such as the NZ Hub to enable their members—and anyone who can adapt their code globally—to take fuller advantage of their membership in ORCID.³⁸

The aim of these national initiatives is to achieve economies of scale, while still allowing for flexibility in how ORCID iDs are implemented across the organization. More recently, as in the UK and Australia, many ORCID projects have been tied into broader PID endeavors.³⁹

Ultimately, Canadian ORCID uptake continues to grow substantially from year-to-year, individually and institutionally, and a number of broader PIDs initiatives including the newly formed CPIDAC will likely continue to emphasize the importance of an open, community-driven, machine-readable PID for people such as ORCID. ORCID iDs are important PIDs in the effort to make research outputs FAIR, and are critical to achieving the potential of a comprehensive linked ecosystem of these outputs.

3.2 Research Outcome Object Identifiers

For many in the research ecosystem, the final published outputs, or the *research outcome objects*, are the primary representation of the findings of a specific research project. This typically means the research publication (or scholarly/journal article), and the dataset(s) associated with that publication. While both types of outcomes use similar PID options, we have divided this section into the two outcomes to highlight important differences and use cases.

3.2.1 Publications

3.2.1.1 Background

Object Identifiers (OIDs) are (typically alphanumeric) strings that establish a unique identity for a specific object. In *Section 3.4, Equipment Identifiers*, *3.5 Physical Sample Identifiers*, and *3.6, Other Identifiers*, we highlight a number of ID systems for describing physical objects of interest to research, but in this case we are interested in specific types of *digital objects*, which can be defined as *a machine-independent data structure consisting of one or more elements in digital form that can be parsed by different information systems; the structure helps to enable interoperability among diverse information systems in the Internet*.⁴⁰ In the research data context the digital objects of interest can be any number of research outputs created during the data

³⁶ NZ ORCID Consortium Members. Royalsociety.org.nz.

<https://www.royalsociety.org.nz/orcid-in-new-zealand/new-zealand-orcid-consortium/who-is-involved-with-the-new-zealand-orcid-consortium/nz-orcid-consortium-members/>

³⁷ ORCID in New Zealand FAQ. Royalsociety.org.nz.

2022. <https://www.royalsociety.org.nz/orcid-in-new-zealand/new-zealand-orcid-consortium/orcid-in-nz-faq/>

³⁸ New Zealand ORCID Hub. Royalsociety.org.nz.

<https://www.royalsociety.org.nz/orcid-in-new-zealand/new-zealand-orcid-consortium/who-is-involved-with-the-new-zealand-orcid-consortium/new-zealand-orcid-hub/>

³⁹ Alice Meadows. *ORCID and the UK National PID Consortium*. ORCID.org, August 3, 2020.

<https://info.orcid.org/orcid-and-the-uk-national-pid-consortium/>

⁴⁰ *Digital Object*. CODATA.org. <https://codata.org/rdm-glossary/digital-object/>

lifecycle, but especially a final output such as a journal article or conference presentation, the dataset associated with such an article or project, or a research data management plan. For our purposes we are interested in an OID system that: 1) uniquely identifies an object; 2) includes a description of that object; and 3) provides a context with which to find that object and its description (as ORCID IDs do for RIDs). This means that an OID will provide an actionable, interoperable, persistent link to a digital object. Once assigned to an item, an OID remains a constant locator, not changing even if an object moves location. While information about a digital object may change over time, including where to find it, the OID for that object will stay constant. When properly maintained, OIDs help solve the problem of dead links and link rot.

Just as with RIDs, OIDs are critical to uniquely identifying research outputs in a landscape where such outputs are created at an ever-increasing pace. When an OID is associated with one or more RIDs the resulting network of associations can present a very powerful way of describing scholars and their outputs.

There are a number of IDs that are being used to uniquely label individual digital objects: the GUID and ARK systems are two examples. A GUIDs (globally unique identifier) is *a unique reference number used as an identifier in computer software*.⁴¹ The GUID (also referred to as the *UUID*, or universally unique identifier) can be used to provide the unique string or label for an object, but in and of itself does not meet our complete needs for an effective OID system. The ARK (Archival Resource Key) is *a multi-purpose persistent identifier for information objects of any type*.⁴² The ARK system was developed by the California Digital Library to provide PIDs for digital objects being created as part of large digitization efforts, but is much more flexible and designed to provide IDs for any type of object, digital, physical, living or intangible.⁴³

The ARK scheme does provide a resolvable OID for research outputs, but ARKs have not been widely adopted in the research landscape to describe the types of research outputs highlighted earlier. Having said that, the use of ARK is increasing for other use cases and it does provide a granular ID system that accommodates the full range of objects. In that broader context, ARKs are a system that should be considered as part of the broader landscape of digital objects, which includes research outputs as a subset. Take the example of a journal article; the article itself can be identified using a widely accepted identifier like a DOI (see below), but the individual *digital components* of the article could also be identified with a PID like an ARK. A standard scientific article has a number of components, including data sets, images, tables, and graphs, and in the case of online journals may also have multimedia elements such as videos or animations. In an ideal scenario each of these individual digital objects would have a unique PID that provides access to that individual component outside the context of the article. In the case of a typical journal system, this level of granularity is not supported, as articles are often represented as a single digital object, such as a PDF file. As journal systems evolve, the opportunity to represent individual components will become more important and systems like

⁴¹ *Universally Unique Identifier*. Wikipedia.org. https://en.wikipedia.org/wiki/Universally_unique_identifier

⁴² *Archival Resource Key*. Wikipedia.org. https://en.wikipedia.org/wiki/Archival_Resource_Key

⁴³ *About ARKs*. ARKS.org. <https://arks.org/about/>

ARK may well provide the best example of a robust PID for those components. An added advantage of ARKs is that the same URL is used for the metadata of the object and the digital object itself. Also, some organizations like ANDS (Australian National Data Service), recommend the use of ARKs with research outputs,⁴⁴ and the creation of the ARK Alliance may also facilitate the use of ARKs.⁴⁵

The primary OID for publications by far is the DOI, which is described both in the introduction to Section 2, and below in other contexts. When it comes to publications, the primary DOI registration agency for English and many other languages is CrossRef, which represents a consortium of over 4,000 publishers that use the DOI system. CrossRef provides DOI services for a variety of object types, including journals, books, conference proceedings, working papers, technical reports, and data sets. There are other registration agencies for publication DOIs, such as the China National Knowledge Infrastructure (CNKI) and the Institute of Scientific and Technical Information of China (ISTIC), the Japan Link Center (JaLC), and the Korea Institute of Science and Technology Information (KISTI). The Multilingual European DOI Registration Agency (mEDRA) is a DOI agency for multilingual outputs in the European context.⁴⁶

The DOI system ensures that every publication is assigned a unique identifier that does not change, even when the location of the object changes. An exception to this is when a journal changes publishers, the new publisher may choose to change the DOIs for all previous issues, although this is not always the case. This approach to persistence is achieved by keeping any changeable attributes of the object in the associated metadata record, which itself is based on the Indecs Content Model or Framework.⁴⁷

3.2.1.2 Technical Description

CrossRef is a not-for-profit organization run by the Publishers International Linking Association Inc. (PILA)⁴⁸, and the metadata associated with that object is openly available via the Initiative for Open Citations (I4OC), which was created to ensure the “unrestricted availability of scholarly citation data”.⁴⁹ CrossRef creates DOIs on behalf of its members, who assign those DOIs to their unique content. Crossref also provides a number of services to facilitate the creation and use of DOIs, including APIs and web interfaces. An example of a Crossref service that may not be obvious to researchers as they peruse the literature, is Reference Linking, which provides links to other research listed in an article’s bibliography, which are added by the publisher. Support for Reference Linking by CrossRef members is mandatory, thus helping to create a rich network of linked scholarly data.

⁴⁴ *Digital Object Identifier (DOI) system for research data*. ANDS.org.au, February 6, 2018.

https://www.ands.org.au/_data/assets/pdf_file/0006/715155/Digital-Object-Identifiers.pdf

⁴⁵ ARK Alliance. ARKS.org. <https://arks.org/>

⁴⁶ DOI Registration Agencies. DOI.org. https://www.doi.org/registration_agencies.html

⁴⁷ *Indecs Content Model*. Wikipedia.org. https://en.wikipedia.org/wiki/Indecs_Content_Model#cite_note-2

⁴⁸ *Membership Terms*. CrossRef.org. http://www.crossref.org/02publishers/59pub_rules.html

⁴⁹ *Initiative for Open Citations*. I4OC.org. <https://i4oc.org>

In a similar fashion, DOIs can be used to provide persistent links to other objects within an individual article, such as tables, graphs, images, etc. This is not as common as providing links to referenced articles, but can be used to create a rich network of linked objects in the scholarly ecosystem. One of the reasons why complex DOIs for elements within a specific article are not common is the cost: every CrossRef minted DOI carries a cost, and that could add up for publishers with complex content. There is also no hard and fast rule for the creation of DOI suffixes, other than the recommendation for “opaque” identifiers, which means the DOI should be a meaningless number, and not contain characters intended to make it human-readable.⁵⁰ Meaningful strings (e.g. dates, chapter or figure numbers) should be part of the metadata record instead.

The examples below are of URLs/DOIs for a single journal article from the Facets journal from Canadian Science Publishing.

1. Article - <https://doi.org/10.1139/facets-2021-0139>
2. Article section - <https://doi.org/10.1139/facets-2021-0139#sec-4>
3. Article figure - <https://doi.org/10.1139/facets-2021-0139#f1>
4. Article table - <https://doi.org/10.1139/facets-2021-0139#tab1>
5. Article supplementary material - <https://doi.org/10.1139/facets-2021-0139#sm1>
6. List of individual article citation in text - <https://www.facetsjournal.com/doi/10.1139/facets-2021-0139#core-ref6>
7. Individual article citation in text - <https://www.facetsjournal.com/doi/10.1139/facets-2021-0139#body-ref-ref6-1>

All of these URLs are resolved properly by the DOI platform, except for the last two, which simply takes you to the top of the article, indicating that the linking functionality at this level is specific to the journal platform (ie. overlapping window structure). The last two allow the user to: view a list of places in the article where a specific external reference is cited; and to highlight a specific occurrence of the cited reference in the text.

3.2.1.3 Use Cases

Some typical use cases for using object identifiers for publications are:

1. A researcher wants to publish in a journal that automatically assigns a DOI, and sends the metadata for the article to the researcher’s ORID profile, facilitating additional promotion of the article.
2. A researcher adds their post-peer review preprint to their institutional repository which provides a unique DOI, providing an open access copy in addition to the paywalled published version.

⁵⁰ *Constructing your DOIs*. CrossRef.org.

<https://www.crossref.org/documentation/member-setup/constructing-your-dois/#afewrules>

3. A publisher joins CrossRef to leverage the services of that DOI authority for its articles, and promotes data repositories that provide complimentary services for data deposit.
4. A funder receives the DOIs for the final outputs of a funded research project, allowing it to more easily follow-up on its open access mandate.

3.2.1.4 Implementation and Impact

The assignment of DOIs for articles is increasingly a standard feature for scholarly journal publishers, providing researchers with an easy route to a more sustainable scholarly record. This applies equally for commercial publishers and open access publishers. Platforms like PKP have good support for DOIs, making it a seamless part of the article submission process.⁵¹ Many institutional repository systems are also capable of integrating with DOI minting services, and some like DSpace, support multiple PID systems.⁵²

Canadian funders and government agencies are increasingly supportive of best practices in open science, including support for the FAIR Principles, which are explicit about the use of PIDs. Many Canadian funding agencies have released data management policies, and the Office of the Chief Science Advisor of Canada released the Roadmap for Open Science in 2020, and it highlights that:

Scientific research outputs are “Open by Design and by Default”; they are “FAIR”, i.e. Findable, Accessible, Interoperable and Reusable. Withholding scientific research outputs requires a valid reason consistent with a framework (to be developed) on which scientific information will be kept private or confidential.

More specifically, the Roadmap makes the following recommendation:

Federal departments and agencies should develop strategies and tools to implement FAIR data principles to ensure interoperability of scientific and research data and metadata standards by January 2023, with a phased plan for full implementation by January 2025.*

A consistent approach to the use of PIDs in this context would be a necessary and critical part of the strategy. Together with the leadership role played by Canada’s higher education institutions, funders and government agencies are well positioned to facilitate access to the scholarly literature.

⁵¹ *Using DOIs and the DOI Plugin*. PKP.SFU.ca <https://docs.pkp.sfu.ca/doi-plugin/en/>

⁵² *DOI Digital Object Identifier*. Lyrasis.org.
<https://wiki.lyrasis.org/display/DSDOC6x/DOI+Digital+Object+Identifier>

3.2.2 Datasets

3.2.2.1 Background

As datasets are increasingly recognized as research outputs in their own right, persistent identifiers are needed for discovery and citation of research data. The FAIR principles articulate the need for globally unique and persistent identifiers for research data (F1):

Globally unique and persistent identifiers remove ambiguity in the meaning of your published data by assigning a unique identifier to every element of metadata and every concept/measurement in your dataset. In this context, identifiers consist of an internet link (e.g., a URL that resolves to a web page that defines the concept such as a particular human protein). Many data repositories will automatically generate globally unique and persistent identifiers to deposited datasets. Identifiers can help other people understand exactly what you mean, and they allow computers to interpret your data in a meaningful way (i.e., computers that are searching for your data or trying to automatically integrate them). Identifiers are essential to the human-machine interoperation that is key to the vision of Open Science and many journals currently require authors to publish their data with their research⁵³. In addition, identifiers will help others to properly cite your work when reusing your data.⁵⁴

Principle F1 applies to both the data itself and its metadata. A dataset should have a persistent identifier assigned. In practice, DOIs have emerged as the most popular identifier for datasets, although several domain specific dataset identifiers are also in common use. In addition, the metadata for a dataset should incorporate other persistent identifiers as appropriate to reference related publications, datasets, people, organizations, funders, software, and other entities.

3.2.2.2 Technical Description

The primary type of identifier used for datasets is a DOI. DataCite is a DOI registration agency with a specific focus on DOIs for datasets. Although DataCite DOIs can be created for any type of resource, the DataCite metadata schema is specifically designed with datasets in mind.⁵⁵ DataCite is a not-for-profit, international organization that works with data centres and researchers to assign DOIs to datasets and other research objects. The aim is to develop an infrastructure that supports data citation, discovery, and access. DataCite does not allocate DOIs itself; this is done by its members, who act as allocating agents. DataCite's DOI minting

⁵³ Reporting standards and availability of data, materials, code and protocols. Nature.com.

<https://www.nature.com/nature-research/editorial-policies/reporting-standards>

⁵⁴ F1: (Meta) Data Are Assigned Globally Unique and Persistent Identifiers. GO-FAIR.org.

<https://www.go-fair.org/fair-principles/f1-meta-data-assigned-globally-unique-persistent-identifiers/>

⁵⁵ DataCite Schema. DataCite.org. <https://schema.datacite.org/> Accessed December 15, 2020.

services include a web interface for minting DOIs—DataCite Fabrica⁵⁶—and several APIs⁵⁷, which have been integrated into different repository platforms.⁵⁸

Other object identifiers that can be used for datasets include Archival Resource Keys (ARKs)⁵⁹, Handles (which may be used independently from the DOI system), URN:NBN⁶⁰, as well as domain-specific PIDs such as accession numbers from national organizations such as the US National Center for Biotechnology Information⁶¹ and the European Bioinformatics Institute.

3.2.2.3 Use Cases

Some typical use cases for using object identifiers for datasets are:

5. A researcher would like to cite the data in their paper so that others can access the data and verify their conclusions.
6. A repository would like to assign a PID to data deposits so that they can be referenced and linked to related material.
7. A publisher wants to promote scientific reproducibility in their journals and therefore requires authors to provide a data availability statement, as well as a citation to where a minimal data set to reproduce the research is stored.
8. A dataset contributor would like to receive credit for the data that they produce, and therefore wants a mechanism for others to cite their data.
9. A community of practice would like to develop a domain-specific discovery and data access layer, using well-described APIs and other technologies to harvest metadata for articles and datasets.

3.2.2.4 Implementation and Impact

In Canada, institutions can create DOIs for datasets and other research outputs via the DataCite Canada Consortium. The DataCite Canada Consortium is a partnership between the Canadian Research Knowledge Network (CRKN) and the Digital Research Alliance of Canada (initiated by the Canadian Association of Research Libraries (CARL) Portage Network) to provide consortial

⁵⁶ *Fabrica Guide*. DataCite.org. <https://support.datacite.org/docs/doi-fabrica> Accessed December 15, 2020.

⁵⁷ *DataCite REST API Guide*. DataCite Support. <https://support.datacite.org/docs/api> Accessed December 15, 2020.

⁵⁸ *Service Provider Software Integrations*. DataCite.org.

<https://support.datacite.org/docs/service-provider-software-integrations> Accessed February 24, 2021

⁵⁹ John Kunze, *ARK Identifiers FAQ*. LYRISIS Wiki. <https://wiki.lyrasis.org/display/ARKs/ARK+Identifiers+FAQ> Accessed September 29, 2020

⁶⁰ Frances Madden et al., *Guides to Choosing Persistent Identifiers - Version 3*. May 28, 2020.

<https://doi.org/10.5281/zenodo.4192174>

⁶¹ National Center for Biotechnology Information. *SRA Metadata and Submission Overview*. NCBI.NLM.NIH.org, accessed March 10, 2022. <https://www.ncbi.nlm.nih.gov/sra/docs/submitmeta/>

membership in DataCite.⁶² The consortial model benefits Canadian institutions by providing lower costs than direct membership, along with additional support and collaboration at the national level. Over 600,000 DOIs have been minted in Canada with DataCite, and over 38,855 DOIs in Canada were minted in 2021.

Prior to the DataCite Canada Consortium, many Canadian institutions minted DOIs through the National Research Council of Canada's DataCite Canada membership. This agreement ended at the end of 2019, and the transition of DataCite Canada members to the new DataCite Canada Consortium took place over the course of 2020. As of February 2021 the DataCite Canada Consortium had 49 members with 59 repositories between them.

The Research Data Alliance Working Group on Dynamic Data Citation published a set of 14 recommendations⁶³ for repositories to support dynamic data and dataset subset citation. Dynamic datasets refer to those that are changing over time as new records are added, errors are corrected, or contents are enhanced. Dataset subsets refer to the notion that researchers often query and use dataset subsets, for which more specific citation is helpful to increase the reproducibility of the research outputs. In the last few years since the recommendations were published, pilot adoptions have demonstrated the viability and broad applicability of these guidelines across varied disciplines and digital infrastructures.

In an effort to have more consistent approaches to data citation text, guidelines such as the ESIP Data Citation Guidelines for Earth Science Data Version 2⁶⁴ have emerged. These guidelines include core concepts such as the creator, public release date, title, repository and persistent identifier, as well as more challenging aspects like dynamic and micro (or subset) citation. The guidance even extends to resolvability and landing page features for the dataset.

3.2.3 Research Software

3.2.3.1 Background

Research software can be described as the software (programs, libraries, scripts, and tools) used to support research. The Open Science, FAIR Data, and reproducibility in science movements have transitioned research software from being a tool that facilitates research to becoming a fundamental part of the research ecosystem. The Digital Research Alliance of Canada released their Research

⁶² Canadian Research Knowledge Network. *DataCite Canada Consortium*. CRKN-RCDR.ca. <https://www.crkn-rcdr.ca/en/datacite-canada-consortium>. Accessed December 15, 2020

⁶³ Andreas Rauber, Ari Asmi, Dieter van Uytvanck, and Stefan Proell. *Identification of Reproducible Subsets for Data Citation, Sharing and Re-Use*. Bulletin of the IEEE Technical Committee on Digital Libraries, 12(1), May 2016. https://bulletin.jcdl.org/Bulletin/v12n1/papers/IEEE-TCDL-DC-2016_paper_1.pdf

⁶⁴ ESIP Data Preservation and Stewardship Committee. *Data Citation Guidelines for Earth Science Data*. Ver. 2. Earth Science Information Partners, 2019. <https://doi.org/10.6084/m9.figshare.8441816>

Software Current State Paper in 2022, providing a detailed description of research software, as well as actors and services in the Canadian context.⁶⁵

Research software plays a triple role in today's research ecosystem⁶⁶:

1. it serves as a tool in many areas by effectively processing various types of data to build and test models to support or invalidate hypotheses;
2. it can be a research result in its own right acting as evidence of an effective algorithmic solution to a given problem as measured by the capabilities of the computers of the day;
3. it can itself be a research object. The research community is particularly interested in the modes of software development and the proof of their properties, especially regarding societal issues related to transparency and trust in computerized processing.

Such software is important at several levels throughout the research lifecycle⁶⁷ and in order to facilitate points 2 and 3 above, persistent identifiers for research software are becoming increasingly important.

It is important to note that there is a difference between the need for citable research software to ensure that the creators of that software *receive credit for their contributions* to research (point 2 above), and the need for citable research for scientific *reproducibility* (point 3 above). In general, publication of a software package in some form (e.g. through a journal publication or through Zenodo) is sufficient for providing academic credit. Both would result in a DOI for the software and therefore provide a mechanism for citation. Scientific reproducibility requires a more fine-grained resolution, as to truly describe a reproducible software pipeline it is necessary to be able to cite a specific version of a software package. As such, PIDs that support scientific reproducibility need to handle changes to the object over time in much the same way PIDs for dynamic datasets do.

3.2.3.2 Technical Description

The FORCE11 Working Group on Software Citation analyzed a number of use cases around software citation, and all of them identified the need for a citable software object.⁶⁸ This group identified DOIs as one of the most prominent mechanisms for creating a PID for a software object, typically through publishing the object through a platform such as Zenodo or figshare. Indeed, GitHub has a software DOI minting service in conjunction with Zenodo.⁶⁹ This practice is

⁶⁵ *Introducing the Current State Papers*. Alliance.ca. <https://alliancecan.ca/latest/introducing-the-current-state-papers> Accessed March 10, 2022.

⁶⁶ Roberto Di Cosmo and François Pellegrini. *Opportunity Note: Encouraging a wider usage of software derived from research*. November, 2019. <https://www.ovvirlascience.fr/opportunity-note-encouraging-a-wider-usage-of-software-derived-from-research/>

⁶⁷ Victoria Stodden, and Sheila Miguez. *Best Practices for Computational Science: Software Infrastructure and Environments for Reproducibility and Extensible Research*. *Journal of Open Research Software* 2 no. 1 (2014): 1-6. <http://dx.doi.org/10.5334/jors.vy>

⁶⁸ Arfon M. Smith et al. *Software citation principles*. *Peer J Computer Science*, 2:e86, 2016. <https://doi.org/10.7717/peerj-cs.86>

⁶⁹ Making Your Code Citable. *GitHub Guides*. <https://guides.github.com/activities/citable-code/>

still evolving, but the primary tools to support the creation of software PIDs do exist. It is becoming more common for research publications to cite PIDs for both data and the research software that was used to produce or analyze that data. For example, Cell Press uses a standard format (Star Methods⁷⁰) to encourage citations to a range of entities including the software used in a publication.

3.2.3.3 Use Cases

The *Use cases and identifier schemes for persistent software source code identification*⁷¹ report produced by FORCE11 and the Research Data Alliance provides a number of specific use cases for all actors in the research software ecosystem, and is a useful guide to what approach should be considered.

- A researcher has developed some software that is useful to a specific research community, and would like to be credited when that software is used by that research community.
- A researcher wants to improve the scientific reproducibility of their research by citing the specific version of a specific piece of software, along with an image of the compute environment, used to perform the analysis carried out in their research.
- A publisher wants to improve the scientific reproducibility of the research published in their journals and therefore requires all publications to cite the software used in the analysis carried out in the publication.
- A funder wants to recognize the value of research software as a research output and requires research projects funded by its funding programs to cite both the software used by research projects as well as cite any research software developed through the research projects.

3.2.3.4 Implementation and Impact

PIDs for software are a relatively new area and are rapidly evolving. There are two fundamental types of software PIDs: extrinsic and intrinsic. Extrinsic identifiers are more familiar, as they involve an external registry for the object. DOIs for software (using Git and Zenodo as described above) are extrinsic because they are created as reference for an object. There are a range of extrinsic identifiers in addition to DOIs in use currently, including ARKs, ASCL-ID, HAL-ID, swMath-ID, RRID, and Wikidata entities⁷². Intrinsic identifiers are derived from the object itself, and are therefore more directly associated with that object. For example, the Software Heritage

⁷⁰ Cell Press: *Star Methods*. Cell Press. <https://www.cell.com/star-authors-guide>

⁷¹ Research Data Alliance/FORCE11 Software Source Code Identification WG, Allen, A., Bandrowski, A., Chan, P., Di Cosmo, R., Fenner, M., Garcia, L., Gruenpeter, M., Jones, C. M., Katz, D. S., Kunze, J., Schubotz, M. & Todorov, I. T. (2020). Use cases and identifier schemes <https://doi.org/10.15497/RDA00053>

⁷² Research Data Alliance/FORCE11 Software Source Code Identification WG, Alice Allen et al. 2020. *Use cases and identifier schemes for persistent software source code identification (V1.1)*. Research Data Alliance. <https://doi.org/10.15497/RDA00053>

project⁷³ has defined a mechanism for generating intrinsic identifiers for software from a variety of research software source code revision management packages.

Software PIDs can be used to refer to research software at a wide variety of granularity, ranging from software projects, project versions, modules, software repositories, releases, commits, files, and even code fragments, depending on the use case.

Currently there are other PIDs used to refer to research software, and there is no clear leading approach and the landscape is rapidly evolving. Both the Research Data Alliance Software Source Code Interest Group⁷⁴ and the FAIR 4 Research Software Working Group^{75 76} are excellent resources to try and understand the current state in this area.

3.3 Organization Identifiers

3.3.1 Background

An organization identifier establishes a unique identity for a specific organizational entity. In the research context, organizations include universities, government agencies, research centres, funders, subunits associated with those entities, and more. Organization identifiers are important for a variety of stakeholders, including academic administrators, funders, publishers, repository managers, software developers, rights agencies and individual researchers. Without organization identifiers, identifying and tracking organizations is challenging because an organization may be known by a variety of names. Organizations may also have several affiliated departments or research centers that are recognized entities in themselves. Organization identifiers can provide the means to both find and identify an organization accurately and to define the relationships with its sub-units.

Organization identifiers can be used for different functions in a variety of use cases. The most common purpose is for tracking *affiliation*: an author's affiliated institution. In some cases, an organization can also be an *author* in its own right, and the organization identifier can be used as a name identifier. Organizations can also function as *funders*. While the organization identifiers covered in this section can be used for funders, there are also specific funding identifiers which will be covered in the following section.

There are several organization identifiers currently in use in the scholarly community, including the Research Organization Registry (ROR), Global Research Identifier Database (GRID), International Standard Name Identifier (ISNI), Ringgold, and Wikidata. Although the scholarly

⁷³ *Software Heritage*. SoftwareHeritage.org. <https://www.softwareheritage.org/>

⁷⁴ Software Source Code IG. RD-Alliance.org. <https://www.rd-alliance.org/groups/software-source-code-ig>

⁷⁵ *FAIR 4 Research Software (FAIR4RS) Working Group*. FORCE11.org. <https://force11.org/groups/fair-4-research-software-fair4rs-working-group/>

⁷⁶ FAIR for Research Software (FAIR4RS) WG. RD-Alliance.org. <https://www.rd-alliance.org/groups/fair-4-research-software-fair4rs-wg>

community is coalescing around ROR, all of these identifiers are still actively used in different systems.

3.3.2 Technical Description

3.3.2.1 Research Organization Registry (ROR)

The Research Organization Registry (ROR) provides identifiers for research organizations. Officially launched in 2019, ROR is “a community-led project to develop an open, sustainable, usable, and unique identifier for every research organization in the world.”⁷⁷ ROR emerged out of several years of collaboration between the California Digital Library, DataCite, Crossref, and ORCID through the Org ID initiative.⁷⁸ Currently, the project is led by the California Digital Library, DataCite, and Crossref, with input from a steering group and a community advisory group.⁷⁹

ROR focuses on the “affiliation use case,” where affiliation is defined as “any formal relationship between a researcher and an organization associated with researchers.”⁸⁰ For example, a faculty member is affiliated with the university where they work, and a student is affiliated with the institution where they study. As of February 2021, the registry currently contains over 99,000 research organizations, over 3,000 of which are for Canadian organizations. The registry is made available via the web interface, an open API, and a data dump, and all data in ROR is CC0 (public domain).⁸¹

The initial data for ROR was donated by Digital Science from the GRID database.⁸² After a period of incubation, Digital Science announced (July 2021) it was discontinuing its GRID releases, and ROR became a fully community-driven organization.

*Since its launch in 2019, ROR’s operating organizations have been working to shore up resourcing and infrastructure. ROR has installed a community-based advisory group and steering group, secured grant funding and community donations, created a governance structure and sustainability model, implemented basic community-based curation workflows, and began building the necessary infrastructure to be able to deploy registry updates independently.*⁸³

ROR aims to be interoperable with different organization identifiers and PID systems. ROR synthesizes multiple organization identifiers in one place: it includes links back to records in the

⁷⁷ Research Organization Registry. *About*. ROR.org. <https://ror.org/about/> Accessed December 8, 2020.

⁷⁸ John Chodacki et al., *Org ID: A Recap and a Hint of Things to Come*. DataCite Blog, August 2, 2018. <https://doi.org/10.5438/67SJ-4Y05>

⁷⁹ Research Organization Registry. *Facts*. ROR.org. <https://ror.org/facts/> Accessed December 8, 2020.

⁸⁰ Research Organization Registry. *Scope*. ROR.org. <https://ror.org/scope/> Accessed December 8, 2020.

⁸¹ Research Organization Registry. *Scope*. ROR.org. <https://ror.org/scope/> Accessed December 8, 2020.

⁸² Research Organization Registry, *Facts*. ROR.org. <https://ror.org/facts/> Accessed December 8, 2020.

⁸³ <https://ror.org/blog/2021-07-12-ror-grid-the-way-forward/>

Crossref Funder Registry, Ringgold, and Wikidata. ROR IDs can be used within the DataCite metadata schema and are integrated in the Fabrica DOI registration system, which can suggest ROR identifiers based on the text entered for an affiliation.^{84,85} There are a growing number of repositories that are integrating ROR identifiers into their metadata—including Ocean Networks Canada⁸⁶ and Dryad⁸⁷ — which take advantage of the updated DataCite schema to add ROR identifiers to affiliation metadata. Additionally, Crossref⁸⁸ and ORCID⁸⁹ are working towards integrating ROR into their systems.

Although ROR is still relatively new, it has achieved international support and has endorsements from the persistent identifier community, librarians, publishers, research organizations, among others.⁹⁰ As of December 2020, ROR is listed by the Library of Congress as one of the Standard Identifier Source Codes.⁹¹ Future plans for ROR include developing the infrastructure to be independent from GRID, expanding the metadata schema to include relationships between research organizations, and establishing community-based curation workflows.

3.3.2.2 GRID

The Global Research Identifier Database (GRID) is a product created by Digital Science. Launched in 2015, GRID's coverage has grown from just over 50,000 research institutions to over 99,000 as of February 2021.⁹² Although GRID is created by a private technology company and lacks community-led governance, the database is openly available in the public domain (initially through a CC-BY 4.0 license⁹³).

Like ROR, GRID is focused on “research-related organizations” and strives for interoperability between different organization identifiers.⁹⁴ Records in GRID contain links to ISNI, the Crossref

⁸⁴ Robin Dasler and Madeleine de Smaele. *Identify Your Affiliation with Metadata Schema 4*. DataCite Blog, August 16, 2019, <https://doi.org/10.5438/vgaq-ar22>

⁸⁵ DataCite. *Fabrica Guide: Field Descriptions for Form*. DataCite.org. <https://support.datacite.org/docs/field-descriptions-for-form> Accessed December 8, 2020.

⁸⁶ Reyna Jenkyns. *Data Partnerships Home*. OcenNetworks.ca, August 18, 2020, <https://wiki.oceannetworks.ca/display/DataPartners#DataPartnershipsHome-ResearchOrganizationRegistry>

⁸⁷ Maria Gould and Daniella Lowenbreg. *ROR-Ing Together: Implementing Organization IDs in Dryad*. ROR Blog, July 10, 2019. <https://ror.org/blog/2019-07-10-ror-ing-together-with-dryad/>

⁸⁸ Rachael Lammey. *Solutions for Identification Problems: A Look at the Research Organization Registry*. Science Editing, 7 no. 1, February 20, 2020: 65–69. <https://doi.org/10.6087/kcse.192>

⁸⁹ Laure Haak. *What Is up with ORCID and ROR?* ORCID Blog, March 25, 2020. <https://orcid.org/blog/2020/03/24/what-orcid-and-ror>

⁹⁰ Research Organization Registry. *Supporters*. ROR.org. <https://ror.org/supporters/> Accessed December 8, 2020

⁹¹ Library of Congress, Network Development & MARC Standards Office. *Source Codes for Vocabularies, Rules, and Schemes: Standard Identifier Source Codes*. LOC.gov. <https://www.loc.gov/standards/sourcelist/standard-identifier.html> Accessed December 8, 2020

⁹² Cameron Shepherd. *Digital Science Launches GRID, a New, Global, Open Database Offering Unique Information on Research Organisations*. Digital Science Blog, October 12, 2015, <https://www.digital-science.com/blog/news/digital-science-launches-grid-a-new-global-open-database-offering-unique-information-on-research-organisations/>

⁹³ Geoffrey Bilder et al. *Organisation Identifiers: Current Provider Survey*. 2016. <https://info.orcid.org/wp-content/uploads/2021/01/20161031-OrqIDProviderSurvey.pdf>

⁹⁴ GRID. *GRID - Global Research Identifier Database*. GRID.ac. <https://grid.ac/> Accessed December 8, 2020.

Funder Registry, Wikidata, Wikipedia, and ROR, as well as the (now retired) OrgRef IDs.⁹⁵ The GRID metadata schema also accommodates geographic coordinates and links to GeoNames IDs.⁹⁶

In addition to providing the seed data for ROR, GRID is also used in other platforms and PID systems. For example, GRID is currently one of the organization identifier types used by ORCID⁹⁷; it is used in several of Digital Science's own projects, including FigShare and Altmetric,⁹⁸ and it is used by publishers including Springer Nature and Hindawi.⁹⁹

3.3.2.3 ISNI

One of the most commonly used organizational IDs with international scope is ISNI (International Standard Naming Identifier), an ISO (International Standards Organization) standard that is used widely and can be assigned to individuals and organizations. The ISNI database currently holds public records of 1,600,000 organizations and over 13 million individuals.¹⁰⁰

3.3.2.4 Ringgold

Ringgold, a for-profit company and publisher-oriented initiative,¹⁰¹ maintains an organizational identifier called the Ringgold ID (or RIN), originally developed to identify institutional subscribers to academic journals. The list of Ringgold IDs is proprietary and is not in its entirety publicly available. Although non-proprietary organizational IDs are growing and evolving, Ringgold IDs remain in use in the open data ecosystem (for example, Ringgold IDs are used, along with GRID and Fundref, in the ORCID Registry to authoritatively identify organizations). Notably, Ringgold was the first "ISNI Registration Agency" for organizational identifiers in 2012.¹⁰²

3.3.3 Use Cases

Some typical use cases for organizational identifiers are:

1. A funder would like to collate all publications related to their funding for a given year.
2. A researcher would like to compile a list of institutions participating in a proposal with multiple partners.

⁹⁵ Nick Andrews. *Retiring OrgRef*. The DataSalon Blog, March 27, 2018.

<https://blog.datasalon.com/2018/03/27/retiring-orgref/>

⁹⁶ GRID, *Policies*. GRID.ac. <https://grid.ac/pages/policies> Accessed December 8, 2020

⁹⁷ ORCID. *How Are Organizations Identified in ORCID?*. ORCID, August 19, 2020.

<https://support.orcid.org/hc/en-us/articles/360006973513-How-are-organizations-identified-in-ORCID->

⁹⁸ Digital Science. *GRID*. DigitalScience.com. <https://www.digital-science.com/products/grid/> Accessed December 8, 2020.

⁹⁹ GRID. *GRID - Global Research Identifier Database*. GRID.ac. <https://grid.ac/> Accessed December 8, 2020.

¹⁰⁰ ISNI. *About ISNI*. [ISNI.org](https://isni.org).

¹⁰¹ NISO. *Institutional Identifiers*. NISO.org. <http://www.niso.org/standards-committees/institutional-identifiers>

¹⁰² <https://www.ringgold.com/isni/>

3. A university administrator would like to identify all publications by their affiliated researchers to assess impact.
4. A repository would like to download all articles published by institutional researchers to ensure that they are complying with an open access policy.

3.3.4 Implementation and Impact

In the past several years, the scholarly community has begun to coalesce around the Research Organization Registry (ROR) as an emerging solution for organization identifiers. A recent Jisc blog post profiled ROR and ISNI, the main two organization identifiers as explored by the UK PID Consortium's focus group on organizations.¹⁰³ The focus group found that "ROR's governance model had greater flexibility and was more likely to be able to respond to emerging needs in the scholarly infrastructure ecosystem."¹⁰⁴ The integration or planned integration of ROR into other PID systems—including DataCite, Crossref, and ORCID—further supports that convergence around ROR is growing.

The FREYA project's *Guide to Choosing Persistent Identifiers* section on Organization Identifiers provides a flow chart for choosing a persistent identifier.¹⁰⁵ For the "research affiliation" use case, ROR, GRID, and Ringgold are the three possible outcomes. ROR is distinguished from GRID by having a "community-led governance structure", while Ringgold is distinguished from ROR and GRID due to not being free of charge.¹⁰⁶ Both ISNI and Ringgold are highlighted for the use case "To identify current and historical public identities (names) of current and defunct organizations," for which the guide's authors do not recommend ROR or GRID.¹⁰⁷ Another difference between ROR and ISNI is the level of granularity: unlike ISNI, ROR does not support department-level affiliations.¹⁰⁸ However, there is work being proposed to extend ROR to include department-level units by including "qualifying tails" on ROR IDs for subunits.¹⁰⁹

¹⁰³ Fiona Murphy and Phill Jones. *There's A PID For That, Part 3: Organisations*. Jisc Scholarly Communications Blog. October 14, 2020.

<https://scholarlycommunications.jiscinvolve.org/wp/2020/10/14/theres-a-pid-for-that-part-3-organisations/>

¹⁰⁴ Fiona Murphy and Phill Jones. *There's A PID For That, Part 3: Organisations*. Jisc Scholarly Communications Blog, October 14, 2020.

<https://scholarlycommunications.jiscinvolve.org/wp/2020/10/14/theres-a-pid-for-that-part-3-organisations/>

¹⁰⁵ Frances Madden et al. *Guides to Choosing Persistent Identifiers - Version 3*. May 28, 2020.

<https://doi.org/10.5281/zenodo.4192174>

¹⁰⁶ Frances Madden et al. *Guides to Choosing Persistent Identifiers - Version 3*. May 28, 2020.

<https://doi.org/10.5281/zenodo.4192174>

¹⁰⁷ Frances Madden et al. *Guides to Choosing Persistent Identifiers - Version 3*. May 28, 2020.

<https://doi.org/10.5281/zenodo.4192174>

¹⁰⁸ Alice Meadows. *Are You Ready to ROR? An Inside Look at This New Organization Identifier Registry*. The Scholarly Kitchen Blog, December 4, 2019.

<https://scholarlykitchen.sspnet.org/2019/12/04/are-you-ready-to-ror-an-inside-look-at-this-new-organization-identifier-registry/>

¹⁰⁹ Liz Krznarich et al. *PIDapalooza 2021: Extending ROR - Wag the Lion*. January 27, 2021.

<https://pidapalooza2021.sched.com/event/gD02/extending-ror-wag-the-lion>

In the Canadian context, language is another important facet impacting the utility of organization identifiers. ROR supports multiple languages, which is essential for a bilingual country.¹¹⁰ However, ROR uses the English name of an institution as the primary “name” field whenever it is available, even if it is not the primary name by which the institution is known. This policy of preferring the English-language name of an organization does not reflect the reality of many Canadian institutions. For example, l’Université de Montréal appears in ROR as “University of Montreal”, which is not the official institution name; “Université de Montréal” appears as both a “label” (associated with the iso639 language code “fr”) and an alias (with no associated language code).¹¹¹ Ideally, Canadian users could retrieve the “official” or “preferred” name(s) of an institution from the ROR API, along with an indication of its language, even if it is not in English. The ROR schema also does not have a means of accounting for institutions with two official names; one of the two names (typically the English name) is used for the primary “name” field, with the other (typically the French name) falling to a label. Furthermore, these issues surrounding language have the potential to impact Indigenous organizations in Canada, whose names should be listed in the language of their choice—even if an English translation is available. With the upcoming move away from GRID’s infrastructure and the GRID metadata schema, we hope that ROR will work to update their schema and curation procedures to better support multilingualism in practice.

3.4 Funding Identifier

3.4.1 Background

The international funder community has embraced the use of PIDs to help ensure research outputs are accessible, but the use of PIDs for grants or other forms of funding has been a gap until more recently. CrossRef and a consortium of international funders launched a DOI registration service in 2019, and it is available for funders now.¹¹² The community is still in development, but the service and memberships are in place. Funders who become members are encouraged to start with new grant programs, and then move to previous programs, as these would be an especially rich source of associated outputs, demonstrating the value of the PID graph.

3.4.2 Technical Description

Since the CrossRef grant PID uses the same DOI framework used with publications and other digital objects, the foundations are very similar. The only way to create CrossRef DOIs currently is by submitting an XML document with the necessary metadata, and there is no online form submission. Funding award types listed in the Funding Type element include: award, contract, crowdfunding, endowment, equipment, facilities, fellowship, grant, loan, prize, salary-award, secondment, seed-funding, training-grant, other.

¹¹⁰ Research Organization Registry. *Facts*. ROR.org. <https://ror.org/facts/> Accessed December 8, 2020.

¹¹¹ *University of Montreal ROR Entry*. ROR.org. <https://ror.org/0161xgx34> Accessed March 10, 2022.

¹¹² CrossRef. *Introduction to grants*. CrossRef.org. <https://www.crossref.org/documentation/content-registration/content-types-intro/grants/>.

3.4.3 Use Cases

1. A funder wishes to have better control of information about their grants over the history of the organization, and also to facilitate the promotion and propagation of the impact of their funding.
2. A researcher would like to refer to all the grant sources in their latest article, and so in a consistent way that facilitates promotion of their research.
3. A research organization would like to reduce the burden of adding all grants received by their researchers to their local research information system.
4. A research infrastructure provider would like to determine which researchers, as well as associated funders, have benefitted from access to their facilities.
5. A publisher would like to have a standard way to automatically add funding information to the DOI record for submitted articles, and to standardize the way funding information is reported in the article text itself.

3.4.4 Implementation and Impact

In 2019, the Treasury Board launched the Guidelines on the Reporting of Grants and Contributions Awards¹¹³, which mandates the public disclosure of grants and awards over \$25,000, and the required and optional metadata is a good fit to that required for a funding DOI. While not requiring the use of a common PID, a Reference Number is a required field, and it is possible that it could be used as part of the DOI suffix, making an implementation even easier. The more challenging aspect of a funding DOI is that the DOI URL should resolve to a landing page for the grant, which is typically not a level of granularity that many funders achieve with their funding calls.

Nonetheless, since most of Canada's federal research agencies must now disclose grants according to the guidelines, it would provide an excellent opportunity to meet this important piece of the PID ecosystem

3.5 Research Project Identifier

3.5.1 Background

The idea of PIDs to uniquely identify specific research projects has been part of some disciplinary frameworks for some time. More recently, the idea of using a standard project identifier across all disciplines has gained momentum, driven by an interest in having a foundational PID that can link all outputs related to a specific project.

¹¹³ Treasury Board of Canada Secretariat. *Guidelines on the Reporting of Grants and Contributions Awards*. TBS-SCT.ca, , accessed March 10, 2022. <https://www.tbs-sct.gc.ca/pol/doc-eng.aspx?id=32563§ion=html>

For example, in the biological sciences, metadata about projects (BioProject¹¹⁴) are captured to encapsulate and link other entities related to a specific study that are stored with accession numbers as PIDs at national or international repositories (such as the NCBI or EBI). The record typically contains basic project information and links to publications and datasets; it may also contain submitter and grant information, links to related publications and relevant external Web resources, the organism name, taxid, and infra-species identifier (strain, breed, cultivar, or isolate).

As in the BioProject example above, projects are frequently described using proprietary IDs defined by local systems, such as research information systems. In funder systems (e.g. Horizon 2020 program) the grant ID is often synonymous with the Project ID. The challenge with this approach is that many research projects (ie. as defined by the researcher) may be associated with multiple grants.

The Research Activity ID (RaID)¹¹⁵ is a recent effort to define a single PID and metadata record that would aggregate PIDs for entities associated with a specific research activity. A research activity is generally described as a project, but a RaID can be any appropriate research activity (e.g. sub-projects, experiments, research programs). The RaID was developed to support research systems in Australia by the Australian Research Data Council (ARDC), and is currently undergoing the ISO process to become an official standard.¹¹⁶

3.5.2 Technical Description

The RaID has two parts: the RaID handle, and the RaID metadata record. The proposed ISO RaID metadata record has a minimal set of required time-date stamped fields, but may contain any number of appropriate PID for entities associated with the project. Like all PIDs, the RaID handle remains the same, while specific associated entities (e.g. researchers, datasets) in the RaID metadata record may change throughout the life of the project.

RaIDs is currently a free service, and RaIDs can be created via the API service for machine-machine integrations (e.g. intended for integration into research information systems and other software platforms), or via the RaID dashboard¹¹⁷ for manual minting services. The goal of the RaID community is to see regional RaID minting services, creating a distributed network of RaID services.

3.5.3 Use Cases

¹¹⁴ National Center for Biotechnology Information. *BioProjects List*. NCBI.NLM.NIH.gov. <https://www.ncbi.nlm.nih.gov/bioproject/browse>

¹¹⁵ *Research Activity Identifier*. RAID.org. <https://www.raid.org.au/>

¹¹⁶ International Standards Organization. *ISO/DIS 23527: Information and documentation — Research activity identifier information technology — Learning, education, training and research (RAiD)*. ISO.org. <https://www.iso.org/standard/75931.html>

¹¹⁷ RAiD Project. *RAiD Dashboard*. RAID.org. <https://www.raid.org.au/dashboard>

- A researcher has carried out a study and would like to create an entity (a “project”), with a PID, that links all of the related information associated with that study, including all team members, associated publications, physical samples gathered as part of the study, and any derived digital data sets that were produced from the analysis of those samples.
- A researcher discovers an article that refers to a number of associated research outputs that are of interest. Clicking on the cited RaID provides a complete listing of all associated entities.
- An infrastructure provider queries a system of federated RaID registries for the ID of specialized equipment funded in previous grant programs, and receives a list of all research projects that used the same equipment.
- Midway through a complex research project, the PI changes, as well as some of the project team members. Previous publications all cited the same RaID, allowing future researchers to see the full history of associated entities, including the current PI and their ORCID record.

3.5.4 Implementation and Impact

RaIDs provide a unique opportunity for the global research community to embrace a common standard for representing the full diversity of research outputs, across all disciplines. By querying RaID associated metadata, and accessing the provided links, innovation can more easily surface related research. An approach like RaID makes the relationship between specific research outputs explicit: without such an explicit record of linked data, creating a picture of the associated research outputs requires what may be termed a “brute force” approach, using often faulty techniques such as extraction of IDs via text analysis of journal articles, which can only produce an incomplete picture. By using a common aggregator of research activity IDs, stakeholders in the research ecosystem would have the ability to traverse a rich international graph of research entities associated with all research efforts.

3.6 Equipment Identifiers

3.6.1 Background

Like the previous examples of IDs, an equipment ID (EID) is a (typically alphanumeric) string that establishes a unique identity for a specific piece of research equipment, or subcomponents of such equipment. Equipment and scientific instruments are increasingly a critical part of many research projects, often generating high volumes of data in a short time. It is important to note that the FAIR Principles “apply not only to ‘data’ in the conventional sense, but also to the algorithms, tools, and workflows that led to that data,”¹¹⁸ emphasizing the value of applying persistent identifiers to equipment used to produce data as well as the dataset itself.

¹¹⁸ Wilkinson, M. D. et al. *The FAIR Guiding Principles for scientific data management and stewardship*. Scientific Data, 3:160018. <https://doi.org/10.1038/sdata.2016.18>

An emerging PID is for instruments¹¹⁹, where instrument is defined as a “device used for making measurements, alone or in conjunction with one or more supplementary devices”¹²⁰. This PID is an outcome of the Research Data Alliance Working Group Persistent Identification of Instruments (PIDINST) since “source information of a dataset (e.g., instrument and method) is essential to interpret the quality of the dataset and to facilitate its reusability, further work should be done to link the remaining and new data submissions with their instrument PIDs, where applicable.”¹²¹

3.6.2 Technical Description

Implementation for the RDA PIDINST instrument PID has been piloted using both DataCite (DOI) and ePIC frameworks. A schema that extends the DataCite kernel has been proposed such that the commonly used metadata can be incorporated. Before being finalized, additional community input is being sought.

3.6.3 Use Cases

Some typical use cases for equipment identifiers are:

1. A funder (e.g. CFI) would like to collate all publications related to the use of a piece of equipment that they helped fund.
2. A researcher would like to compile a list of all collaborators who have used data from a specific piece of equipment that they acquired and managed.
3. A collaborator would like to see the history of actions (e.g. installations, configuration changes, software upgrades) taken with a piece of equipment that generated data used in their study.
4. An equipment manufacturer would like to identify uses of their equipment, including specific examples of research outputs.
5. A university computing centre would like to track citations and usage of their high-performance computing platform. (<http://doi.org/10.14288/SOCKEYE>).

A full set of use cases from the PIDINST WG is available in Github.¹²²

¹¹⁹ Christine Ferguson et al. *D3.1 Survey of Current PID Services Landscape (Version 1)*. Zenodo, July 17 2018. <https://zenodo.org/record/1324296>

¹²⁰ International Vocabulary of Metrology. *Basic and General Concepts and Associated Terms (VIM 3rd edition)*, 2012. https://www.bipm.org/documents/20126/2071204/JCGM_200_2012.pdf/f0e1ad45-d337-bbeb-53a6-15fe649d0ff1?version=1.15&t=1641292389029&download=true

¹²¹ M. Stocker et al. *Persistent Identification of Instruments*. Data Science Journal, 19(1) 2020, p.18. <http://doi.org/10.5334/dsj-2020-018>

¹²² RDA WG PIDINST. *Use Cases*. <https://github.com/rdawg-pidinst/use-cases>

3.6.4 Implementation and Impact

The use of EIDs has only recently started to receive significant attention, so recommendations regarding best practices are only now being developed. However, given the increasing relevance of this component of the research ecosystem, we felt it important to highlight the latest developments. One of the most useful efforts with EIDs is ‘equipment.data’, a project co-funded by Jisc and the Engineering and Physical Sciences Research Council.¹²³ The goal of the project is *to improve visibility and utilisation of UK research equipment*.¹²⁴ In addition to providing a searchable database of research equipment in the UK, the system has defined a metadata record (UNQUIP¹²⁵) that is used to ensure a common record format for the equipment. When institutions submit a record it includes a local ID (ie. an institutionally derived ID for a piece of equipment) and the equipment.data project then mints a unique ID based on an PD5 hash value derived from a specific algorithm.¹²⁶ A sample ID is reflected below.

- <https://equipment.data.ac.uk/item/27e1f9d552d01ff87ff36fe919bfaab1.html>

While the equipment.data project is a good example of the development of an approach to EIDs, it is used in a specific context and does not offer a solution for an international standard to the generation of PIDs and associated URIs for equipment.

Another important initiative in this space is the MERIL project in the EU, which is *the only systematic inventory of significant research infrastructures that is comprehensive over all European countries and scientific domains. MERIL provides high-quality information about research infrastructures and data extraction capabilities on an open-access web portal*.¹²⁷ The MERIL database provides access to information at a higher level than individual pieces of equipment (e.g. facilities or Centres), but the project is starting to consider the issue of individual EIDs.

There have been a few pilot implementations of the PIDINST solution, including by the British Oceanographic Data Centre (BODC) and the Helmholtz-Zentrum Berlin für Materialien und Energie (HZB) that demonstrate the practical viability.¹²⁸ EUDATs PIDs 4 Instruments service is intended to provide a landing page for research infrastructures in the EU.¹²⁹ The working group is now actively working with the community on adoption of their proposed approaches.

¹²³ Equipment Data Project. *Equipment Data Homepage*. Equipment.data.ac.uk. <http://equipment.data.ac.uk>

¹²⁴ Equipment Data Project. *Equipment Data Homepage*. Equipment.data.ac.uk. <http://equipment.data.ac.uk>

¹²⁵ Equipment Data Project. *UNQUIP Data Publishing Specification*. Equipment.data.ac.uk. <http://equipment.data.ac.uk/uniquip>

¹²⁶ Personal correspondence, Adrian Cox, University of Southampton. Jul 29, 2016.

¹²⁷ MERIL- Mapping of the European Infrastructure Landscape. ESF.org <http://www.esf.org/serving-science/ec-contracts-coordination/meril-mapping-of-the-european-research-infrastructure-landscape.html>

¹²⁸ M. Stocker et al. *Persistent Identification of Instruments*. Data Science Journal, 19(1), p.18. <http://doi.org/10.5334/dsj-2020-018>

¹²⁹ EUDAT. *Register and publish your scientific instruments*. <https://b2inst-poc2.eoschub-surfsara.surf-hosted.nl/>

A recent analysis highlighting the PIDINST outcomes asserts that “the application of the FAIR principles to physical aspects of research can greatly improve the preservation, interpretation and reusability of this type of research.”¹³⁰

3.7 Physical Sample Identifiers

3.7.1 Background

In the biological sciences, metadata about physical biological samples (BioSample) that are associated with studies are often stored at national or international repositories such as the NCBI or EBI. These repositories typically use accession numbers as PIDs to link samples to both the more general construct of a research project (BioProject) as well as digital data sets that are derived from these samples using technologies such as next generation sequencing. In the biological sciences, knowing information about the physical sample is critical, because this is the entity that contains key biological metadata about the sample (tissue, disease state) as well as the subject (age, gender, anonymized subject id) from which it was taken. This information is key for both data reuse and scientific reproducibility.

An effort by the research community resulted in the development of the Research Resource ID (RRID),¹³¹ and is intended to provide unique IDs for biological resources, including antibodies, model organisms, cell lines, plasmids, and other Tools (software, databases, services).¹³² The RRID Portal at SciCrunch is the primary RRID portal, using a federated search approach to query a host of appropriate databases and registries. The Biodiversity Information Standards organization (TDWG) recommends the use of GUIDs and LSIDs in the biodiversity community, although use by the broader community is unclear.¹³³

The International Geo Sample Number (IGSN) became a global initiative in 2011 with the founding of the IGSN e.V., a non-profit organization that governs the IGSN. The IGSN is a unique alphanumeric code that is assigned to physical samples such as rocks or sediment cores to ensure their unique identification and unambiguous referencing for the generated data and results. Originally applied to geological samples, the usage has expanded to other types including fluid and biological samples. Many journals in the earth and physical sciences use IGSN's, and they are recommended for use by key community groups such as COPDESS and PANGAEA.

3.7.2 Technical Description

¹³⁰ E. Plomp. *Going Digital: Persistent Identifiers for Research Samples, Resources and Instruments*. Data Science Journal, 19(1), 2020 p.46. <http://doi.org/10.5334/dsj-2020-046>

¹³¹ RRID. *Current Project*. RRID.org. <https://www.rrids.org/current-project>

¹³² SciCrunch. *RRID Portal*. SciCrunch.org. <https://scicrunch.org/resources>

¹³³ TDWG. *GUID and Life Sciences Identifiers Applicability Statements*. TDWG.org. <https://www.tdwg.org/standards/guid-as/>

The IGSN currently supports two types of membership, full and affiliate members. There were 16 full members and 5 affiliate members as of December 2020.¹³⁴ A key feature of full membership is the ability to become an allocating agent, which can operate registration services and issue IGSNs on behalf of the main IGSN registration agency (refer to the full roles and responsibilities is available on the website¹³⁵). The Lamont Doherty Earth Observatory was awarded Sloan Foundation funding for the IGSN 2040 project which was intended to upscale and modernize the overall system.¹³⁶ Improvements to the metadata schema, vocabularies and more will be revised based on outcomes of community consultations and related working group efforts. SESAR is a community-maintained IGSN allocating agent, and is intended to provide an open and accessible platform for the use of IGSNs.¹³⁷

RRIDs are not dissimilar to DOIs in their syntax:

*RRIDs are prefixed with "RRID: ", followed by a second tag that indicates the source authority that provided it (e.g. "AB" for the Antibody Registry, "CVCL" for the Cellosaurus, "RGD" for Rat Genome Database, "SCR" for the scicrunch registry of tools).*¹³⁸

The examples below show URLs for a specific antibody in both the SciCrunch Portal, and the Antibody Registry, both using the same RRID. The value proposition is further illustrated by the integration of RRIDs into an creator's ORCID record.¹³⁹

- https://scicrunch.org/resources/Any/search?q=AB_528484
- https://antibodyregistry.org/search.php?q=AB_528484

3.7.3 Use Cases

- A researcher in the biological sciences wants a persistent identifier (BioSample accession number) for the metadata about the biological materials utilized in their study such that the sample can be linked to the digital sequence data that is derived from that sample.
- A scientist would like to identify the physical sample used in a research publication in order to request for a sub-sample to reproduce or extend upon the results.

3.7.4 Implementation and Impact

¹³⁴ IGSN. *Members*. IGSN.org. <https://igsn.github.io/membership/>

¹³⁵ IGSN. *Allocating Agents*. IGSN.org. <https://www.igsn.org/allocating-agents/>

¹³⁶ Marie Denoia Aronsohn. *Sloan Foundation Grant Will Help Support Open and Transparent Science*. State of the Planet Blog, July 20, 2018. <https://blogs.ei.columbia.edu/2018/07/20/sloan-foundation-grant-open-science/>

¹³⁷ SESAR Project. *Welcome to SESAR*. GeoSamples.org, accessed March 10, 2022. <https://www.geosamples.org/>

¹³⁸ Christine Ferguson et al. *D3.1 Survey of Current PID Services Landscape - Revised*. Zenodo.org. May 31, 2018. <https://zenodo.org/record/3554255>

¹³⁹ Anita bandrowski. *Using ORCID to Connect Researchers and their Antibodies*. ORCID.org Blog, July 29, 2019. <https://info.orcid.org/using-orcid-to-connect-researchers-and-their-antibodies/>

The BioSample repository was created in 2011 by NCBI and as part of the international INSDC collaboration between the US (NIH/NCBI), EU (EBI), and Japan (DDBJ) has widespread adoption in the biological sciences. Currently there are over 16 million biological samples recorded in the BioSample database.

RRIDs are widely used in the biosciences, and are required in over 2,000 journals, so would be considered the current best practice for the biosciences. As illustrated in Figure 1 with the PID ecosystem, in a best practice PID ecosystem, a researcher or other stakeholder could land on any individual PID, such as a RRID or IGSN, and get to any other object associated with that research program.

3.8 Other Identifiers

In the previous sections, we have described IDs that together contribute to a more cohesive and linked scholarly landscape. However, identifiers can also be used in a wide variety of other contexts and practices, providing additional citable components in dissemination of research and further facilitating discoverability. Although not the focus of this report, it is important to note that the generalized use of other IDs in other contexts will create an environment that improves access, discovery, impact and recognition in the research domain. As with researchers, research outcomes (including publications, datasets, and research software), organizations, funders, research projects, equipment, and physical samples, using unique IDs for other entities will facilitate name-disambiguation, knowledge discovery, and ultimately the reuse of valuable scholarly materials.

The list below provides examples of best practice IDs for specific domains of research, as well as general IDs that have not yet become widely used. Domain-specific IDs are numerous (e.g. often internally generated accession numbers unique to each platform), and substantial infrastructure development may be needed to turn a unique ID into a PID. The examples below include both IDs and PIDs. Some of the examples have been extracted from the excellent work of the FREYA Project.¹⁴⁰

Research Domain/Service	ID/PID Name	More Information	Maturity & Examples ¹⁴¹
Domain Examples			
Science	MIRIAM/Identifiers.org, URIs	Wikipedia , Registry	Mature BioAssay
Mycology	MycoBank ID	Wikipedia , MycoBank	Mature Species

¹⁴⁰ Christine Ferguson et al. *D3.1 Survey of Current PID Services Landscape - Revised*. Zenodo.org. May 31, 2018. <https://zenodo.org/record/3554255>

¹⁴¹ Based on the terminology used by the FREYA Project: Mature, Emerging, Immature

Chemistry	InChi	Wikipedia , InChI Trust	Mature Article About the InChi
Pharmaceuticals	DIN	Wikipedia , DPD	Mature Drug
Enzymes	EC Number	Wikipedia	Mature Enzyme Catalyzed Reaction
Proteins	PDB ID	Protein DataBank	Mature COVID-19 Protein
Genomics	BioProject, BioSample, Sequence Reads	International Nucleotide Sequence Database Collaboration	Mature COVID-19 Genome
Domain-Agnostic Examples			
Data Management Plan	DOI, RaID	PIDapalooza 2021 DataCite Blog	Immature DOIs considered for the final output, RAiDs also being explored as viable options.
Conference	DOI, Accession Number	CrossRef Page	Emerging Metadata record agreed to, implementation underway
Repositories	Re3data ID, DOI	Scientific Data Article Re3data FAIRsharing	Emerging re3data records use an internal ID, FAIRSharing records use an internal ID and a DOI .
Analyses	Git gists, CERN CAPs		Immature
Workflows	UUID, DOI	MyExperiment	Immature
Computer Simulation	UUID, DOI	CERN OpenData	Emerging
Protocols	DOI	Protocols.io	Immature

Clinical Trials	Accession #, DOI	ISCRTN	Immature
-----------------	---------------------	------------------------	----------

3.9 The Future of PIDs

Like all technology, the systems and frameworks that underlie the PID ecosystem are subject to change, which can be especially problematic for a technology that is intended to identify and point to a digital object “in perpetuity”. While the creators of PID systems clearly consider the issue of persistence, the success of these implementations in practice can be spotty. The reasons for this could vary, but could include: the disappearance of an agency (or central authority) that had sole oversight for a specific PID system; a change in the business model of a PID authority, causing previous users to move to a different system or abandon it; or variations in citation practices for digital objects that have multiple URIs.

Some of these challenges are accommodated by infrastructure providers by developing a technology migration policy, specifically to deal with some of the longer term gaps associated with changes in technology, policy, funding, etc.

This section describes one recent effort to redefine the approach to a PID ecosystem.

DIDs

Decentralized identifiers (DIDs) is a new and promising technology for creating identifiers currently and is not yet implemented in the research space. Existing identifiers typically require a centralized registry which all parties in the ecosystem need to trust, but this centralized model is prone to security breaches and can be hard to scale. Centralization also creates gatekeepers controlling the flow of users and information, which may limit the potential of research identifiers. Unlike more traditional research identifiers, DIDs require no centralized authority to create identifiers as cryptographic technologies ensure each created identifier is unique (duplications or collisions¹⁴² are so improbable they can be ignored).

The DID core specification is currently being standardized at the World Wide Web Consortium (W3C)¹⁴³ and describes new ways of creating, managing, verifying, resolving, and retiring these persistent decentralized identifiers. Implementation details such as the choice of verifiable data registry system (e.g. blockchain, distributed ledger technologies, distributed file systems, peer-to-peer networks) depend on the specific DID method (currently 110+ and counting).

Each DID identifies a single resource or subject and resolves to exactly one DID document. DID documents include the cryptographic signature of the controller(s) providing built-in verifiability, and additional metadata about the subject, such as external endpoints which connect the DID

¹⁴² Reed, D., Preukschat, A. (2020). Self-Sovereign Identity. v8. Meaning Publications.

¹⁴³ W3C Organization. *Decentralized Identifiers (DIDs) v1.0*. <https://www.w3.org/TR/did-core/>

identifier to resources such as webpages, other DIDs or semantic descriptions. The DID specification also allows statements of other identifiers the subject is known as, for example the DOI and ARK identifiers describing the same journal article. Multiple endpoints means that the subject of the DID can itself consist of multiple objects, like a project which can contain authors, datasets, publications, institutions and departments each identified and described by their own DID and all listed as endpoints within the project DID document. Through continual updating of the DID document, research resources can have an evolving permanence where information is added and subtracted as the research identifier landscape changes while the DID remains unchanged.

There is considerable opportunity in developing DIDs applied to research identifiers and open development is currently occurring within the Trust over IP Foundation¹⁴⁴.

4. Discussion

Research is increasingly international and multidisciplinary, resulting in a very complex research environment. Unique identifiers facilitate discovery and reuse of content, global interoperability, and a better understanding of the impact and value of scholarship. They also reduce time and administrative burden by enabling the entry of data into one system that can be automatically reused in the context of other systems, supporting more streamlined and automated processes for researchers, funders, vendors and institutions. They also facilitate the transfer of information across platforms and organizations, and establish links across systems institutionally, nationally and internationally. Used together, RIDs, research outcome identifiers, organization identifiers, RaIDs, and other PIDs represent a sophisticated infrastructure of interconnected people, organizations and resources that enable the community to innovate in new ways. Clearly there are many advantages to adopting unique identifiers for research entities and outputs.

However, to be useful, identifiers must be more than just distinctive strings of alphanumeric characters; they should be integrated and compliant with the various infrastructures and systems that use them. They should also be both human and machine readable and actionable, ideally with API access, resolvability and a recognisable, open license such as CC0 (Creative Commons Public Domain). When identifiers are standard and machine readable, it facilitates the synchronization of disparate data events via “triggers”: systems can automatically update each other every time a new record is added, or an existing record is changed. For example, DataCite can automatically push an update from DOI metadata to the author’s ORCID profile using this kind of actionable metadata infrastructure.¹⁴⁵ The opportunity this presents to ease both the researcher’s and the organization’s administrative burden is substantial, and critical to ensuring that research outputs are accessible. If more local research information systems

¹⁴⁴ Trust Over IP Foundation. *Trust Over IP Website*. <https://trustoverip.org/>

¹⁴⁵ Martin Fenner. *Explaining the DataCite/ORCID Auto-update*. DataCite Blog, October 2019, 2015. <https://blog.datacite.org/explaining-the-datacite-orcid-auto-update/>

supported a standard like RaID, the opportunity to document all appropriate research outputs associated with specific projects would represent a substantial leap in scholarly communications, and making research outputs FAIR.

Despite their obvious advantages, there have been some concerns expressed about the widespread use of PIDs. For example, there is a lack of privacy in a scholarly environment which relies heavily on PIDs. Once a PID is associated with a person or object related to a person, that information is in the public sphere forever. While the scholarly community has always relied on personal identifiers, such as student IDs or email addresses, in the past this information has not been made available to the public, and organizations make significant efforts to ensure that this information remains confidential. It is worth recognizing that once PID services are widely adopted, there will be no way to remove this information from the internet. On the other hand, most scholarly outputs are already part of the public record through their existence in open and proprietary databases, and many would argue that the outputs of publicly funded research must be known.

Another issue is that PIDs may be vulnerable to failure, commercialization, or a change in mission. It is important, therefore, that the solutions we adopt adhere to principles and governance that ensure services are developed based on the needs of the community and universally accepted best practices for the governance of open standards. As such, PID services should be open, freely available, easily accessible, and maintained in a well-governed, independent, trusted and sustainable manner.

The ODIN project (ORCID and DataCite Interoperability Network) defined the term “trusted identifier” as part of a conceptual model of PID interoperability. In this model, a trusted identifier is one that is *unique*, *persistent/resolvable*, *descriptive/discoverable*, *interoperable* and *governed*. These characteristics are further defined¹⁴⁶:

- **Unique** identifiers are unique on a global scale, allowing large numbers of unique identifiers
- **Persistent** identifiers resolve as HTTP URIs with support for content negotiation, and these HTTP URIs should be persistent.
- **Descriptive** identifiers come with metadata that describe their most relevant properties, including a minimum set of common metadata elements. A search of metadata elements across all trusted identifiers of that service should be possible [to aid discovery].
- **Interoperable** identifiers are interoperable with other identifiers through metadata elements that describe their relationship.

¹⁴⁶ Martin Fenner et al. *ODIN: the ORCID and DataCite interoperability network*. International Journal of Knowledge and Learning, 9 (4) May 2015. <https://doi.org/10.1504/IJKL.2014.069537>

- **Governed** identifiers are issued and managed by an organization that focuses on that goal as its primary mission, has a sustainable business model and a critical mass of member organizations that have agreed to common procedures and policies, has a governing body, and is committed to using open technologies.

Additionally, not all persistent identifiers are made equal. A robust PID system should adhere to best practices including¹⁴⁷:

Actionable: They should be resolvable on the web.

Syntax: The syntactic qualities of the PID should be flexible and scalable.

Supporting Services, Interoperability, Community:

- The PID service should not come bundled together with other services nor create technical or administrative dependencies
- The PID service should be reliable, sustainable, secure, and cost effective and there should be no administrative or technical obstacles to using the services
- The PID services should use a formal and well-documented standard, flexible enough to interoperate with other schemes, and not dependent on protocols which may change over time or become obsolete
- The PID should be mature, well-supported, and widely-adopted system with a committed community of users
- The PID service should enable privacy controls so that individuals can set their own levels of access

The following is an idealized use case from the Biological Sciences that links many of the PID concepts discussed above. Note that the papers referenced and the PID citation use case is real, but not all of the processes or PIDs are used in this fashion today.

A research group publishes a paper on the Adaptive Immune Response (AIRR) in patients with COVID-19 (e.g. Schultheiß et al.¹⁴⁸), publishing in a journal that provides a DOI for their paper. This is the key PID for the research output for this study. The authors use a software PID (DOI) to reference the software they use to process the data and an equipment PID (another DOI) to identify the sequencing equipment used to produce the sequence data. In the journal publication, the researcher might refer to a number of other related entities using PIDs such as the funding the authors received to do the research and the funding agencies that provided it (Crossref Funder ID). In addition, the authors or the journal itself might use PIDs to identify the authors (ORCID), PIDs for the institutions to which they are affiliated (ROR), as well as PIDs for the research groups that contributed to the paper.

¹⁴⁷ *Persistent identification and PIDs*. Social History Portal.

http://hopewiki.socialhistoryportal.org/index.php/Persistent_Identification_and_PIDs

¹⁴⁸ Schultheiß et al. *Next-Generation Sequencing of T and B Cell Receptor Repertoires from COVID-19 Patients Showed Signatures Associated with Severity of Disease*. *Immunity*, Volume 53, Issue 2, P442-455.E4, August 18, 2020. <https://doi.org/10.1016/j.immuni.2020.06.024>

The research group generates a large sequencing dataset about the immune response of the subjects in their paper (AIRR-seq data). As good citizens of the Open Science community and to facilitate research producibility, the authors want to share their data. At the same time, the journal requires the publication of related data as part of its Open Science and data sharing policy. As such, they need a PID for their dataset. The researcher uploads the basic sequence data to the National Centre for Biotechnology Information (NCBI) or the European Bioinformatics Institute (EBI). NCBI and EBI provide mechanisms to describe metadata about the project (with accession numbers as PIDs), including entities that describe the Project (BioProject), the Physical Sample (BioSample) as well as the sequence read data (the dataset) itself. The digital sequence read dataset produced by the study is loaded into the NCBI Sequence Read Archive (SRA) or the EBI European Nucleotide Archive (ENA). All of these entities are linked within NCBI/EBI such that one can find all of the sequence data from a specific project. Each sequence read data set from a specific individual of a specific type (e.g. B cell or T cell) is loaded as a separate data set with a separate accession number as a PID. The Adaptive Immune Response Repertoire (AIRR) Community has established a standard (the MiAIRR standard¹⁴⁹) for storing AIRR-seq data and a process for recording this data in NCBI.

The researcher uses the NCBI/EBI BioProject accession number as the key PID to link data from their study to their paper. If necessary, the researchers can refer to the sequence read data from a specific sample with an accession number in their paper as well.

To further increase the value of the data, the research group processes the sequence read data as deposited in the ENA and produces a secondary dataset with annotations at the sequence level to help identify specific B-cell and T-cell receptors. This data is then stored in a repository in the AIRR Data Commons¹⁵⁰ (identified by a repository PID), making the annotated sequences discussed in the paper more readily re-usable by other researchers exploring the immune response to COVID-19. This processed dataset is assigned a PID such that the processed data can be referenced specifically. The AIRR Data Commons, and specifically the MiAIRR standard recommends that researchers use NCBI/EBI accession numbers to link datasets in the AIRR Data Commons to their source sequence data in repositories like NCBI/EBI. Data in the AIRR Data Commons has links to other external PIDs such as the DOI for papers, ORCIDs for researchers, as well as the accession numbers for the data and metadata in the NCBI/EBI repositories.

This processed study data might then be reused and referenced by other authors (e.g. Heming et al.¹⁵¹), including use of the specific dataset from this study. The referring paper references the

¹⁴⁹ Rubelt et al. *Adaptive Immune Receptor Repertoire Community recommendations for sharing immune-repertoire sequencing data*. *Nature Immunology*. 18, pages 1274–1278(2017) <https://dx.doi.org/10.1038%2Fni.3873>

¹⁵⁰ Christley et al. *The ADC API: A Web API for the Programmatic Query of the AIRR Data Commons*. *Frontiers in Big Data*. 17 June 2020, <https://doi.org/10.3389/fdata.2020.00022>

¹⁵¹ Heming et al. *Neurological Manifestations of COVID-19 Feature T Cell Exhaustion and Dedifferentiated Monocytes in Cerebrospinal Fluid*, *Immunity*, Volume 54, Issue 1, P164-175.E6, January 12, 2021. <https://doi.org/10.1016/j.immuni.2020.12.011>

paper (citation/DOI), the PIDs for the repository the data was extracted from, and the PID for the specific processed datasets that were reused.

4.1 Governance and Sustainability

As with many aspects of digital research infrastructure, the challenge of building sustainable long-term frameworks and supports is the biggest challenge in most, if not all jurisdictions. The issue of sustainability emerges at all levels:

1. individual researchers do not always have time to understand what needs to be done to properly apply PIDs to their outputs, especially as the technology changes over time;
2. institutions may lack the capacity to support researchers, or funding to support regional or national efforts to sustain a PID framework;
3. national RDM organizations do not always see it as their mandate to provide sustainable support for a common set of approaches across the country;
4. national funding agencies do not always have the mandate or technical capacity needed to effectively understand and deploy resources that can support other levels in the research ecosystem;
5. non-profit or domain-specific research infrastructure providers may lack the capacity to ensure sustainable support for PID services in their platforms, and are often relying on project-based funding;
6. non-profit PID research infrastructure providers require a substantial ramp-up period before PID standards are broadly endorsed, and the community exists to provide appropriate supports to sustain these organizations;
7. for-profit research infrastructure providers may feel their interests, and those of their customers, are best supported by existing proprietary PID services, and lack the capacity or business need to support open standards.

These are just a few of the issues that can cause the PID ecosystem to appear fragmented and unable to effectively embrace the huge potential of a robust PID ecosystem. The Recommendations section below provides suggestions for ways that this situation can be improved.

One recent development intended to respond to the governance and sustainability gap is the development of the Canadian PIDs Advisory Committee (CPIDAC), mentioned earlier in this document. CPIDAC provides a model of multi-institutional, multi-PID agency collaboration to address issues of governance and sustainability. It will be critical for CPIDAC to adopt a model that facilitates the integration of all actors in the PID ecosystem going forward. In 2021 CPIDAC launched an effort to develop a National PIDs roadmap, working with the MoreBrains Cooperative¹⁵² to initiate the planning for this strategy. It is anticipated that the CPIDAC will embark on the next stage of this effort in 2022.

¹⁵² MoreBrains Cooperative. MoreBrains Website. <https://www.morebrainsconsulting.coop>

The CPIDAC strategic effort was supported by the Digital Research Alliance of Canada, which also heard from its stakeholder community during an extensive needs assessment process, that PIDs are an important part of the research ecosystem. PID-focused papers were *Persistent Identifiers in Canada – Position Paper*¹⁵³, *Persistent Identifiers in Canada: ORCID-CA and DataCite Canada*¹⁵⁴, and *The opportunities of Decentralized Resource Identifiers in the research landscape*¹⁵⁵.

Finally, there are efforts globally to develop support for a global PID ecosystem. Examples include interest and working groups in the Research Data Alliance (RDA) and the PIDapalooza community. The April 2021 Plenary meeting of the RDA saw the first meeting of the *National PID Strategies Birds of a Feather* group, which will discuss opportunities to define a more cohesive approach to global governance and sustainability of PID agencies and infrastructures.

4.2 Recommendations

Our goal in updating this report is to support the efforts at creating a responsive and sustainable PID ecosystem in Canada, and to encourage the adoption of PID services that adhere to best practices and the principles of trust and good governance.

The recommendations below have been edited, and some new recommendations have been added to reflect new developments in the ecosystem.

All Stakeholders

General	1. Commit to the adoption of best practice PIDs and ensure that the solutions adopted are not-for-profit, with appropriate community governance.
Researcher IDs	2. Join the ORCID-CA consortium to develop a comprehensive national solution for the adoption of PIDs in the research ecosystem.
Output IDs	3. Join the DataCite Canada Consortium to develop a comprehensive national solution for the adoption of PIDs in the research ecosystem.

¹⁵³ Eugene Barsky. *Persistent Identifiers in Canada – Position Paper*. Alliancecan.ca. <https://alliancecan.ca/white-papers/persistent-identifiers-in-canada-position-paper>

¹⁵⁴ John Aspler. *Persistent Identifiers in Canada: ORCID-CA and DataCite Canada - White Paper*. Alliancecan.ca. <https://alliancecan.ca/assets/uploads/documents/whitepapers/PIDs-White-Paper-Final.pdf>

¹⁵⁵ Carly Huitema et al. *The opportunities of Decentralized Resource Identifiers in the research landscape*. Alliancecan.ca. <https://alliancecan.ca/assets/uploads/documents/whitepapers/The-opportunities-of-Decentralized-Resource-Identifiers-in-the-research-landscape.pdf>

National PID Strategy	4. Participate in the development of a national PID Strategy for Canada.
-----------------------	--

Research Funders

Researcher IDs	<p>5. Require researchers to have an ORCID iD when applying for grants.</p> <p>6. Work with Common CV Office to include the integration of ORCID iDs and the ORCID Repository into the CCV system.</p> <p>7. Invest in platforms/features that allow funding awards to be pushed to recipients ORCID records.</p>
DOIs	8. Require the adoption of DOIs (or an equivalent domain-specific PID) for data sharing and as an outcome of approved Funder data management policies and principles.
Institutional IDs	<p>9. Adopt ROR as the standard for institutional identifiers.</p> <p>9.1 Ensure that the information about all relevant research centres and affiliates are accurately documented in the ROR system to facilitate the integration with local, regional and national systems.</p>

Universities and Research Centres

Researcher IDs	<p>10. Develop institutional policies and strategies that encourage the adoption of ORCIDs, and best practices in describing and citing research outputs.</p> <p>11. Work with CARL, regional library consortia, and the CARL Portage Network on the development and deployment of educational and awareness resources that highlight the value of ORCIDs.</p>
DOIs	12. Require the adoption of DOIs for data sharing in the context of data management policies and principles.
Institutional IDs	13. Adopt ROR as the standard for institutional identifiers.

	13.1 Ensure that the information about all relevant research centres and affiliates are accurately documented in the ROR system to facilitate the integration with local, regional and national systems.
RAiDs	14. Look at opportunities to adopt the emerging RAiD standard to uniquely identify research projects, and provide an PID aggregation service to facilitate discovery.

Science-based Government Departments

Researcher IDs	15 Require researchers to have an ORCID when applying for grants.
DOIs	16. Require the adoption of DOIs for data sharing in the context of data management policies and principles.
Institutional IDs	17. Adopt ROR as the standard for institutional identifiers 17.1 Ensure that the information about all relevant research centres and affiliates are accurately documented in the ROR system to facilitate the integration with local, regional and national systems.

Repositories and Publishers

General	18. Support the adoption of ORCID iDs, DOIs, ROR, Funding, Funder IDs, and other PIDs into metadata records in order to facilitate the development of a linked metadata framework.
---------	--

Researchers

Researcher IDs	19. Register for an ORCID Identifier and complete the ORCID Profile by adding research outputs.
DOIs	20. Deposit datasets and other research outputs with repositories that provide DOIs and use the DOI when citing that dataset or research output.
Other PIDs	21. Ensure that all possible best practice PIDs are used in key outputs, such as journal articles.

Digital Research Alliance of Canada and other National Agencies

General	<p>22. Develop a strategy for the widespread adoption of PIDs, working with other stakeholders to facilitate those developments.</p> <p>22.1 Working with CPIDAC and other appropriate groups, create a <i>PID Action Committee</i> to act on and deploy a strategy for national adoption of PIDs.</p> <p>22.2 Task the <i>PID Action Committee</i> with updating this document on an ongoing basis, and creating an Action Plan, as advancement in PID systems warrant.</p> <p>22.3 Develop a communications strategy to encourage the adoption of PIDs in Canada</p> <p>23. Require the use of PIDs in all outputs from Alliance-funded projects.</p> <p>24. Engage with the international research data management community on the development and adoption of PIDs.</p>
Researcher IDs	<p>25. Continue to support national PID organizations, including CPIDAC, the ORCID-CA consortium, and DataCite Canada Consortium.</p> <p>26. Contact Canadian journal publishers with the goal of encouraging the adoption of a “PID Mandate” policy, requiring the use of ORCIDs, DOIs, ROR, and other best practice PIDs in article/dataset submission.</p>
DOIs	<p>27. Promote the use of DataCite DOIs for research data sets in Canada and engage with developments internationally to ensure best practice in Canada.</p>
RAiDs	<p>28. Review the opportunity to become the national RAiD Authority for Canada, and promote the adoption of RAiDs in the Canadian PID ecosystem.</p>
Other PIDs	<p>29. Promote the adoption of best practice PIDs for all research outputs, including recommendations for how all actors in the research ecosystem can to integrate these PIDs into their workflows.</p>

5. Acknowledgements

We are very grateful to the RDC community for their contribution to this document, but especially:

- Sankarshan Mukhopadhyay and other members of [Trust Over IP](#).