MORE BRAINS

Incentives to Invest in Identifiers

A cost-benefit analysis of persistent identifiers in Australian research systems

AUSTRALIAN ACCESS FEDERATION

ARDC
Australian Research Data Commons

MORE BRAINS

Table of Contents

MORE BRAINS

Authors:
Josh Brown, Phill Jones, Alice Meadows, Fiona Murphy

DOI: 10.5281/zenodo.7100578

MORE BRAINS

Acknowledgements

The authors would like to thank the following people for their time, effort and thoughtful contributions to the research that went into this report:

Hannah Allan, Curtin University
Melroy Almeida, AAF
Amir Aryani, Swinburne University
Carolyn Bray, Federation University Australia
Adrian Burton, ARDC
Claire Carter, University of Wollongong
Antoinette Cass, Bond University
Keely Chapman, RMIT University
Lachlan Charles, TERN
Julie Clift, Curtin University
Tom Demeranville ORCID
Anusuriya Devaraju, TERN
Lisa Dooner, Southern Cross University
Michelle Duryea, Edith Cowan University
Liz Eedle, Universities Australia
Lynette Farquhar, Griffith University
Elleina Filippi, AAF
Qian Garrett, University of Notre Dame, Australia
Elisa Maria Girola, TERN
Mandeep Goraya, ARC
Marianne Gration, Flinders University
Daniel Hook, Digital Science
Simon Huggard, Swinburne University
Clare Job, University of Wollongong
Em Johnson, Swinburne University of Technology
Tony Krizan, NHMRC
Erin Le Nevez, Australian National University
Matthias Liffers, ARDC
Stephanie McGlinchy, Australian Catholic University
Regina Magierowski, University of Tasmania
Heath Marks, AAF
Natalie Mast, Murdoch University
Ian Mitchell, University of Tasmania
Peter Newman, Flinders University

MOREBRAINS

Scott Nicholls, University of Western Australia
Linda O'Brien, Griffith University
Meg O'Connor, Dept of Health, NMHS, WA
Oona O'Gorman, ARC
Margie Pembroke, Southern Cross University
Cel Pilapil, ARDC
Thomas Reeson, Griffith University
Shawn Ross, Macquarie University/ARDC
Joe Shapter, University of Queensland
Natasha Simons, ARDC
Anya Smeaton, Western Sydney University
Amberyn Thomas, University of Queensland
Peter Vats, Research Graph
Matthew Weyland, Monash University
Tania Wilmann, James Cook University
Lyle Winton, The University of Melbourne
Justin Withers, ARC

The authors would like to thank all the ARMS and CAUL members who contributed to the study by participating in online webinars and discussions, and by providing data for the study via surveys.

This paper was written using data obtained on 12th April, 2022, from Digital Science's Dimensions platform, available at https://app.dimensions.ai. Access was granted to subscription-only data sources under licence agreement

Executive Summary

Persistent identifiers (PIDs) are unique alpha-numeric codes that positively identify entities such as people, places, and things. In addition, they are connected to registries of information about those entities, known as metadata, that enable robust linking to and between those entities. This establishes provenance and attribution, as specified by the FAIR data principles[1]. PIDs contribute to research integrity and reproducibility by precisely identifying the resources used to conduct research and the outputs that result from it. The ability to link research activities to their inputs and outputs bolsters research integrity and facilitates the gathering of evidence for improved strategic decision-making at the individual, institutional, and national levels.

The Australian Research Data Commons (ARDC) and the Australian Access Federation (AAF) commissioned the MoreBrains Cooperative to undertake an analysis of the incentives for adoption of persistent identifiers (PIDs) by the Australian research sector. The three main benefits of PIDs are:

- **Metadata reuse**: PID registries act as both repositories for metadata, and as services that can provide programmatic access to it, saving the time and effort of rekeying it, and improving accuracy.
- **Automation**: The presence of a PID in a system or a metadata record can act as a trigger for an action. The value of automation can go beyond time saved to include more complete information and more timely information processing.
- **Aggregation and analysis**: At the institutional or national scale, aggregating information about entities and the relationships between them enables strategic analysis, benchmarking, the plotting of trends, and other insights.

This report sets out the benefits of PIDs primarily through the first of these lenses: metadata reuse. This is the most amenable to quantification, as data is available about the number of specific entities in the Australian research system (such as the number of researchers, institutions, publications, and grants). By combining this information with existing research on metadata use, the time taken for various kinds of manual data entry, and staff costs in Australia, we are able to attach both time and dollar values to the savings that comprehensive PID adoption could bring:

- The total time cost of this tedious work is nearly **38,000 person days per year.**
- The direct financial cost of this wasted effort is nearly **$24 million per year.** Accounting for the opportunity cost associated with technology transfer and innovation-led growth suggests a far higher figure of **$84 million per year.**

Our primary recommendations are as follows:

- Develop a national PID strategy for Australia, which builds on the success of the AAF-led Australian ORCID consortium and leverages the leadership ARDC is already providing on PIDs.
- Key stakeholders in the Australian research sector—such as universities, research institutions, funders, and infrastructure providers—should integrate a suite of five priority PIDs: ORCIDs for people, ROR for institutions, RAiDs for projects, DOIs for research outputs, and DOIs for grants.
- As part of a longer-term strategy, work should continue on developing PIDs for Instruments, expanding the uses of IGSNs for samples, and potentially other IDs, in collaboration with research communities.
- Funders should build on the success of Australian Research Council's (ARC) integration of ORCID into their research management system by adopting a similar approach and expanding to include the full suite of priority PIDs.
- Commercial providers of Research Information Management Systems (RIMS) and repositories and the communities that support open source RIMS should be engaged to encourage and enable the further wholesale adoption of PIDs into those systems.
- Ensure widespread adoption of PID workflows, with a target for 80% adoption of the five priority PIDs within five years

1. Introduction

Persistent identifiers (PIDs) are unique digital codes associated with an entity (such as a person, publication, or a scientific instrument) and linked to descriptive information about it (metadata). PIDs are a critical component of national research infrastructure, providing not only the ability to uniquely identify entities, but also to link them in a way that establishes research provenance and attribution in a rigorous and resilient way, as specified by the FAIR data principles, which stands for Findable, Accessible, Interoperable, Reusable[1]. The ability to link research activities to their inputs and outputs bolsters research integrity and facilitates the gathering of evidence for improved strategic decision-making at the individual, institutional, and national levels.

PIDs offer a way to embed metadata into descriptions and records of entities at the point of creation or publication. In addition, they offer a way to openly store metadata in standardised formats that are both human- and machine-readable, thereby facilitating information exchange and eliminating the need to tediously rekey information into multiple systems.

In globally challenging times, we depend on accurate, timely dissemination of research results and activities more than ever. The same challenges make compelling demands on the public purse, and on those delivering research with public funds to do so as efficiently as possible. These demands are heightened by concerns about ever-increasing time pressures faced by academics. In the recent Statement of Expectations from the Hon Jason Clare MP, Minister for Education, sent to Judi Zeilke, the Chief Executive of the Australian Research Council (ARC) in August of 2022[2], Mr Clare wrote:

> *I note the higher education sector's concern regarding the workload required for the current mode of delivery of the ERA assessment.*

Beyond the Excellence in Research for Australia (ERA) assessment, Mr Clare also referred to the administrative burden taken on by researchers when applying for research grants[3]:

> *Streamlining the processes undertaken during National Competitive Grant Program funding rounds must be a high priority for the ARC … I ask that the ARC identify ways to minimise administrative burden on researchers.*

Concerns about administrative workload have been building over recent years and are the subject of much discussion. In ARC's 2020 review of both the ERA and Engagement and

[1] For more information about the FAIR data principles, see https://www.go-fair.org/fair-principles/

[2] A complete text of the letter can be read here: https://www.arc.gov.au/about-arc/our-organisation/statement-expectations-2022

[3] A case study of the impact of ORCID integration into the ARC's research management system, which is used for competitive grant applications can be found in section 7

Impact Assessment (EI)[2], the existing strengths of Australia's excellent research environment and community were recognised, but several areas for improvement were also identified. In particular, the reduction of administrative burden on universities was seen as a key priority. The focus on reduction of burden to free up researchers' time to conduct meaningful and impactful research in Australia's national interest has therefore become a key policy goal. It is for this reason that the Australian Research Data Commons (ARDC) and Australian Access Federation (AAF) commissioned the MoreBrains Cooperative to investigate how PIDs can reduce that burden and increase efficiency in the Australian academic research sector.

Calculating the proportion of time that researchers spend on administrative tasks rather than research is challenging, but estimates do exist. A study led by the University of Florida estimated the proportion at 42%[3]; Julia Miller at the University of Adelaide estimated it at 35%[4]. These figures are alarming but, anecdotally, researchers and research managers confirm that they are not surprising. The 2018 inquiry into funding Australia's research conducted by the Parliament of Australia heard evidence that support the need for greater efficiency in both pre- and post-award workflows[5]. For example:

> *In many cases variation documentation and processes required by funding agencies are overly onerous, with no net benefit to researcher or funder.*
>
> *-University of Wollongong (Submission 50, pp 6-7)*

2. PID-enabled workflows

There has been increased interest in PIDs as a means to simplify and automate the exchange of information. In an interview with AAF, published on their website in 2021[4], Professor Joe Shapter, Pro-Vice-Chancellor for Research Infrastructure at the University of Queensland, gave a personal account of his experience using the Open Researcher And Contributor Identifier (ORCID). By using his ORCID ID when publishing, and connecting his ARC profile to his ORCID record, his publication record is automatically added to his grant applications, saving three or four days per submission—as he put it, "*a mountainous saving of work*".

Stories like this are the reason why countries around the world are developing national PID strategies, and prioritising investment in these fundamental information infrastructures[6]. In the United Kingdom (UK), work on one such strategy led to a cost-benefit analysis of PID integrations[7]. The potential benefits of PIDs were weighed against the costs of integrating them in digital platforms and services, and of supporting the UK research community in adopting them comprehensively. The study found that significant cost savings would be realised by investing in a suite of PID integrations and that the provision of a national support network to reduce barriers to participation is justified. The resulting improvements in research efficiency from freeing up research time could lead to as much as £420M of benefit to the UK economy per year.

Through the research done in the UK context[8] and elsewhere, five priority PIDs for research entities were identified. These are:

- DOIs for funding grants
- DOIs for outputs (eg publications, datasets, etc)
- ORCIDs for people
- RAiDs for projects
- ROR for research performing organisations

Exploring the ways that these PIDs could serve a range of policy needs, from reduced administrative burden to FAIR data, MoreBrains generated PID-enabled workflows, which identify the touchpoints at which information about people, places, projects, works, etc are entered into systems or retrieved from them. Figure 1*[9]*, gives an overview of how PIDs can be used to facilitate information exchange through central metadata registries.

[4] Joe's story was an interview given by Joe Shapter on Aug. 04, 2021 and published on AAF's blog as an ORCID user story here: https://aaf.edu.au/orcid-user-stories/ (accessed Sep. 21, 2022).

Between 2020 and 2022, MoreBrains led a series of community consultation activities, commissioned by Jisc[5] and sponsored by Research England. This involved consulting 75 researchers, funders, research managers, and librarians from around the world, including representatives from ARDC, AAF, and ARC. This consultation led to the creation of more detailed views of four specific workflows; institutional research management, funding, research data, and publications[10]. These workflows are intended to inspire all those working to support research and innovation to consider how they could use existing powerful, global, open research information infrastructures in new ways to improve the health and resilience of the research ecosystem.

[5] Jisc is a not-for-profit, membership organisation that provides network, IT and digital resources in support of research, further and higher education institutions, as well as for not-for-profits and the public sector. https://www.jisc.ac.uk/

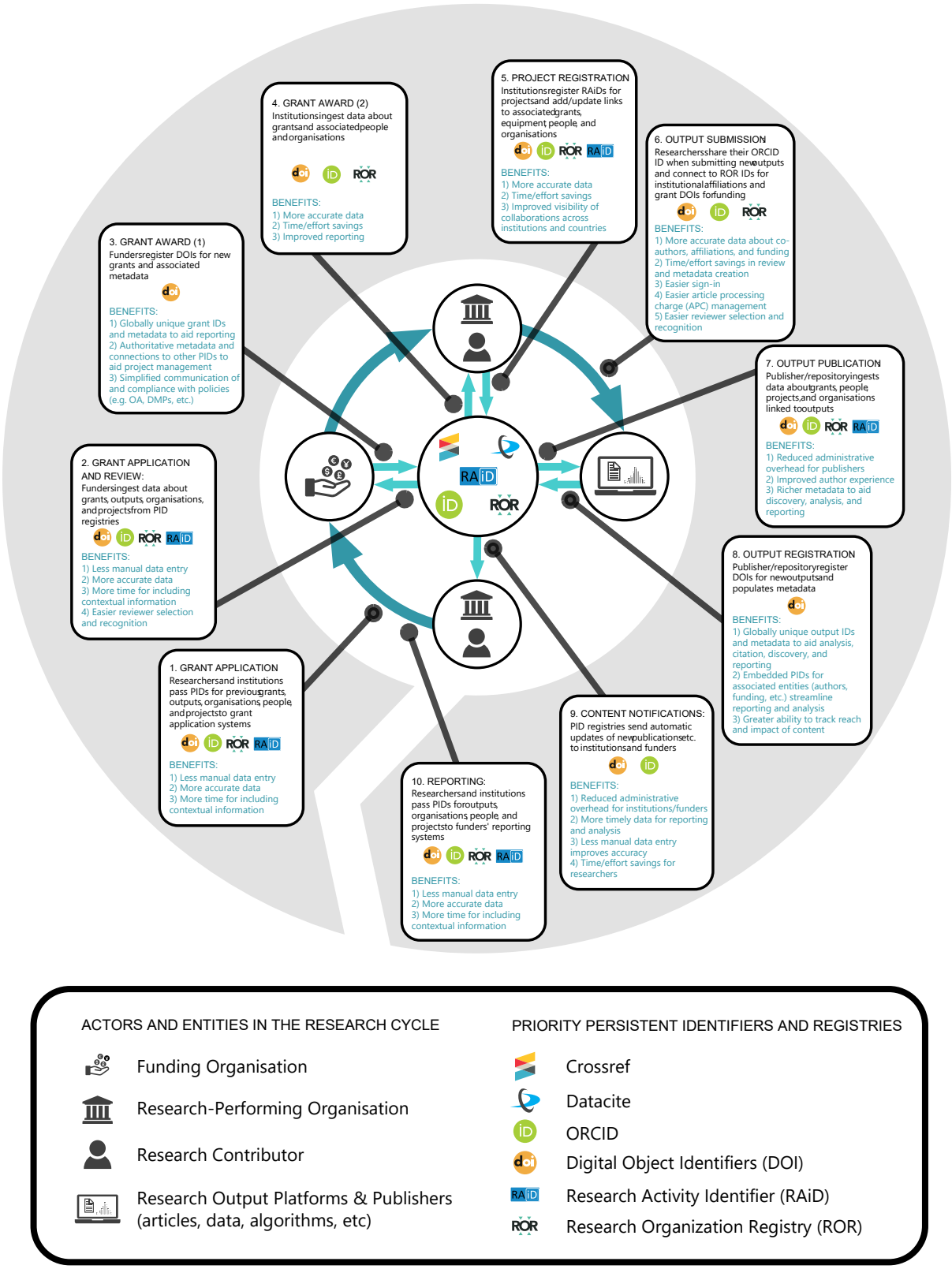

Figure 1: The PID-enabled research lifecycle, developed by MoreBrains, gives a general view of how PID registries can facilitate the transfer of metadata automatically, thereby reducing time and expense. https://resources.morebrains.coop/pidcycle/ (DOI: 10.5281/zenodo.4991733)

3. PIDs in Australia

Australian organisations have been leaders in the support and adoption of PIDs. In 2014, an Australian ORCID Working Group was convened, including representatives from the Australian National Data Service (ANDS), the Council of Australian University libraries (CAUL), the Australasian Research Management Society (ARMS), Australian Research Council (ARC), Universities Australia (UA), and the National Health and Medical Research Council (NHMRC), who drafted a joint statement of principles on unique researcher identification[6]. In April 2015, two joint statements—one by UA, ARMS, CAUL and ANDS[7], the other by ARC and NHMRC[8]—were released in coordination, as a result of the working group's efforts. The ARC/NHMRC statement set out two main mechanisms by which the use of ORCIDs can streamline research administration:

1. Facilitating disambiguation of researchers and research outputs.
2. Enabling the linking and reuse of high-quality, persistent data (e.g. publications, grants).

Building on the joint statements on researcher identification, the Australian ORCID consortium model was developed and collaboratively facilitated by eight organisations: UA, ARC, NHMRC, ARMS, CAUL, Council of Australian University Directors of Information Technology (CAUDIT), ANDS, and AAF. The consortium currently has 43 members: 38 universities, as well as three non-university research-performing organisations and two funders[9]. As detailed in section 9, the consortium runs on a cost recovery basis and has saved the Australian academic research sector in excess of $4.5M, based on calculated savings of $108k per member in licence fees, since its launch in 2016.

ANDS, one of the three organisations that combined to form the ARDC, was a founding member of DataCite, an international Registration Agency for DOIs that can be assigned to research data and related materials. ARDC has continued the ANDS tradition of covering all DataCite fees so that around 70 Australian universities and research institutions can register and manage DOIs for their data collections. ARDC also offers a range of other PID services

---

[6] More details about the joint statements can be found in an ORCID blog post by Natasha Simons, published in April 2015: https://info.orcid.org/orcid-joint-statements-launched-in-australia/

[7] A copy of the UA, ARMs, CAUL, & ANDS joint statement can be found archived via the Wayback Machine https://web.archive.org/web/20200213232325/http://ands.org.au/__data/assets/pdf_file/0007/417346/orcid-joint-statement-of-principle.pdf

[8] The HHMRC & ARC joint statement is available from the ARC website here: https://www.arc.gov.au/news-publications/media/communiques/nhmrc-and-arc-statement-open-researcher-and-contributor-id-orcid

[9] A full list of Australian ORCID consortium members can be found on the AAF website here: https://aaf.edu.au/orcid/members/

on a similar no cost model, reducing both cost and administrative burden to research institutions across Australia,

ARDC also provides Handles, a general purpose identifier frequently used by institutional repositories to identify datasets, collections, publications and a variety of other outputs. Handles are the technology that underpin the Digital Object Identifier (DOI) network and are being used by ARDC as the foundation for the Research Activity Identifier (RAiD), a research project identifier[10] that is under active development by ARDC. It will provide interoperability and an open standard for describing research projects. RAiD is expected to become a certified ISO Standard, and the RAiD service is already in demand internationally, including a project to incorporate RAiDs in the European Open Science Cloud.

Other entities of significant interest to the Australian research community include scientific instruments and physical samples. PIDs for these entities include International Generic Sample Numbers (IGSNs), previously known as the International Geo Sample Number, for physical samples—an important identifier for Australia's world-leading research in geoscience, climate change, and the environment. Section 8 includes information about the planned adoption of IGSNs at the Terrestrial Ecosystem Research Network (TERN). These two categories of PIDs were added to the list of five priority PIDs as candidates for this analysis. The viability of a cost-benefit analysis for each of the seven PIDs is assessed in section 4.

Australia's leadership in the early adoption of many PIDs has been cemented by the creation of high-quality services and robust community governance mechanisms to support the targeted adoption of PIDs like ORCIDs and DataCite DOIs. It is now clear that taking the benefits these have already delivered to the next level requires a new national strategy. The cost-benefit analysis and associated evidence gathered in this report can be used as an input to the development of this strategy, informing a holistic, sector-wide view of a PID-optimised future for Australian research that integrates the efforts of institutions, funders, and technology or system providers alike. This approach will create the environment and conditions needed to maximise the efficiency gains and benefits realisation for existing and new investments in PIDs, and to maintain Australia's leadership in this space.

[10] More details about the RAiD identifier can be found on the RAiD website: https://www.raid.org.au/

4. Scale of research activity in Australia

As the potential benefits of PIDs increase in proportion to the number of entities they identify, we quantified the scale of research activity in Australia by deriving likely numbers of key entities involved in research. These include:

- Universities (acknowledging that there are numerous non-university research-performing organisations, such as disciplinary institutes or hospitals)
- Researchers
- Funding grants
- Publications
- Projects
- Instruments
- Samples

We combined data from various sources to create a plausible upper bound for the number of each entity, and relied on data from PID registries, public records, and analytics platforms to estimate PID coverage for those entities wherever this data was available.

4.1. Institutions

According to the most recent data from the Higher Education Research Data Collection (HERDC)[11], there are 42 "Table A" and "Table B" universities (called providers) as defined in the 2003 Higher Education Support Act (2003)[11], this number does not take account of the range of research institutes and facilities that also operate in Australia.

While all the universities have a PID from the Research Organisation Registry, or the International Standard Name Identifier (ISNI), we have no way of assessing the coverage of organisational PIDs for distinct institutes 'hosted' by a larger organisation or made up of collaborative contributions from multiple partner organisations. The nature of adoption for these identifiers is also distinct from the other entities explored below, as it hinges on the integration of a relatively small number of PIDs (compare the number of universities to the number of researchers employed within them) in a wide array of systems and databases. Further research is needed to assess the actual levels of adoption and usage of these PIDs.

[11] HERDC official data. Available Online: https://app.powerbi.com/view?r=eyJrIjoiZDQxMDZmN2UtNTc3OS00OWU3LThiMWQtYmVlNzgzNDhkYTEyIiwidCI6ImRkMGNmZDE1LTQ1NTgtNGIxMi04YmFkLWVhMjY5ODRmYzQxNyJ9

4.2. Researchers

Again, relying on the most recent data from HERDC, there are 108,873 FTE academic staff employed in universities in Australia.

There is no definitive total for the number of researchers actively using ORCID in Australia. ORCID data is about people so, for data protection reasons, the full dataset cannot be made publicly available. However, ORCID consortium statistics provide a total number of researchers in the ORCID registry, derived from the number of records associated with a .au email suffix, of 171k. In addition, Tom Demeranville of ORCID indicated that of these records, 122k are 'active' (defined as being updated or logged into in the last 365 days at the time of writing in September 2022).

ORCID coverage exceeds the HERDC number as each individual has an ORCID, irrespective of whether they are full- or part-time and it also includes contributors to research such as project managers and research service providers, who are not categorised as 'researchers' by their terms of employment. It is also likely that there are additional records for Australian researchers that are missing from this count as they have not filled in an affiliation, or have used a personal email address to register for their ORCID ID.

4.3. Research inputs (grants)

According to the 2018 ERA, $10.9bn in research income was reported, including competitive research grants from funders like ARC and NHMRC (category 1), public sector R&D income (category 2), industry funding (category 3), and cooperative research centre grants (category 4). Since data on the number of grants and funding sources was only gathered for category 1 grants, we had to look elsewhere for estimates of the number of grants that require metadata entry at Australian universities.

To estimate the number of grants issued to Australian institutions, publicly available data for the yearly funding allocation from ARC[12] was combined with data from the Medical Research Future Fund,[13] and data from the Digital Science Dimensions database[12] [14]. The resulting graph is shown in Figure 2. There is a fairly steady rate of about 6,000 grants awarded each year, with ARC and NHMRC awarding the vast majority.

[12] Data for ARC funding is freely available from the ARC website: https://www.arc.gov.au/grants-and-funding/apply-funding/grants-dataset
[13] Data for MRFF is freely available on the health.gov.au website: https://www.health.gov.au/resources/publications/medical-research-future-fund-mrff-grant-recipients
[14] Digital Science. (2018-) Dimensions [Software] available from https://app.dimensions.ai. Accessed on 12th April, 2022, under licence agreement.

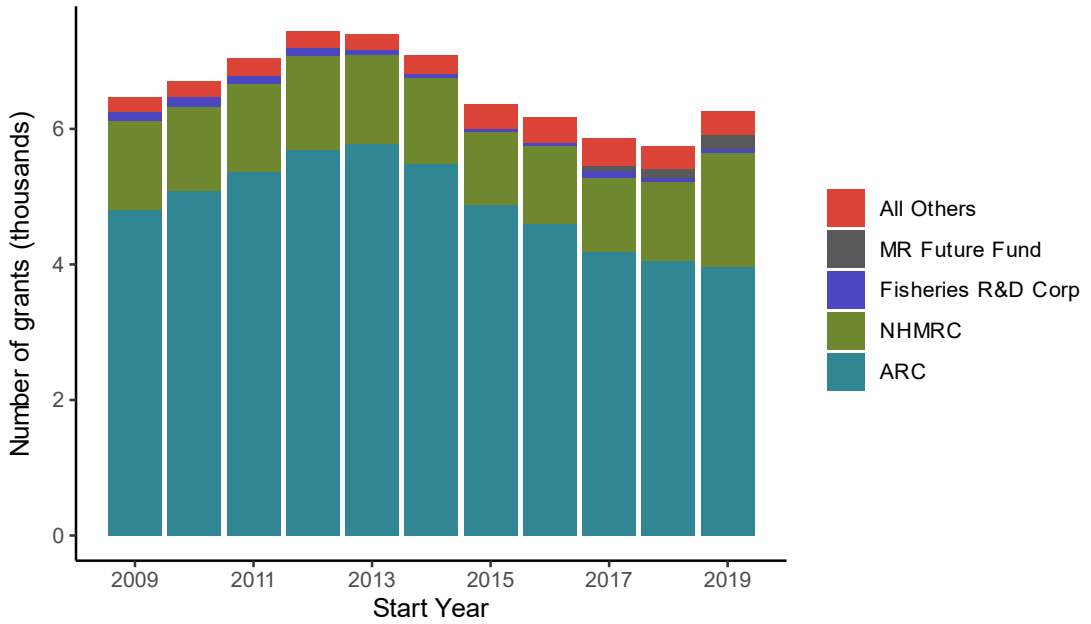

Figure 2:The number of grants awarded to researchers at Australian institutions based on data from ARC, the Medical Research Future Fund, and the Dimensions database

There is a long-standing practice of using Persistent Uniform Resource Locators (PURLs) to uniquely identify grants in Australia. Since their adoption in 2009, ARDC have issued 68,259 PURLs for ARC and NHMRC as of time of writing (September 2022). The PURL system is functionally similar to the resolution services offered by DOIs; in that they consist of a stable URL that redirects the user to a particular web page. Provided the redirect is maintained, the PURL will function to route users to the resource, even if its URL changes. Unlike DOIs, PURLs do not have inherently associated metadata or an agreed schema[13]. In the case of ARDC's PURLs, metadata associated with these grants can be found at a grants portal managed by ARDC[15]; following a PURL takes the user to a human-readable landing page containing information about the grant[16].

DOIs have recently emerged as a best practice for identifying funded grants as a result of work inspired by the Crossref funder advisory group led to the creation of a strategic plan and the launch of their registration service for grants in 2019[17]. The adoption of DOIs for grants is therefore in its early stages. However, at the time of writing, ARDC have issued 77 Crossref DOIs for their own co-investment projects for research infrastructure.

[15] The Grants portal can be accessed here: https://researchdata.edu.au/grants
[16] This URL http://purl.org/au-research/grants/arc/LP0455245 is an example of PURL. In this case, it resolves to a landing page on the ARC data portal
[17] More information on DOIs for grants from Crossref and be found here: https://www.crossref.org/community/grants/

4.4. Research outputs (publications)

Due to challenges around data availability, good estimates of the total research output for Australian universities are difficult to obtain. To estimate the number of publications produced in Australia, we used the Dimensions database again, the data from which is shown in Figure 3. We found that the number of publications from Australian universities is steadily rising from about 74,000 in 2009 to just over 180,000 published in 2019.

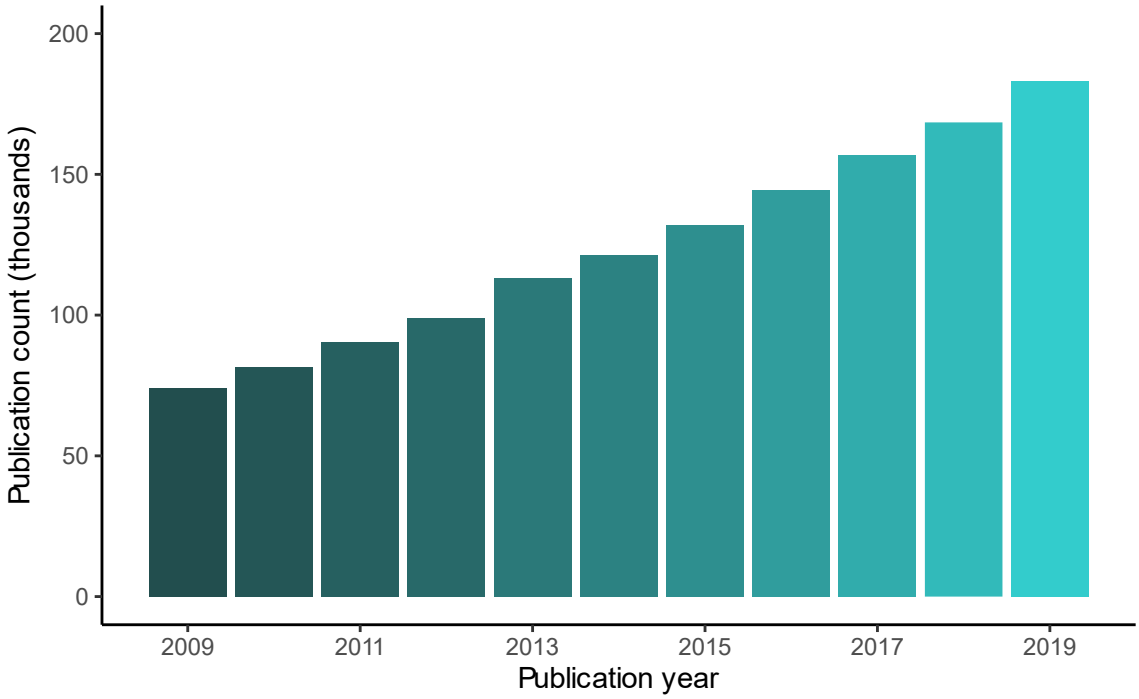

Figure 3: The number of publications in the Digital Science Dimensions database associated with Australian institutions

Of the outputs counted here, there is close to 100% coverage in DOI registries, primarily for articles via Crossref. For other types of output (books, book chapters, reports etc.) it is harder to assess current levels of coverage.

4.5. Projects

Using published numbers for levels of research funding[18], we assumed that the number of projects scales in proportion to the level of funding. Using the UK estimate of approximately 50,000[7] active research projects in universities at any given time, and knowing that Australian research and development funding is equivalent to approximately 42% of funding in the UK, we estimate that there are approximately 21,000 active research projects at any given time in Australian universities.

[18] OECD data is available here: https://data.oecd.org/rd/gross-domestic-spending-on-r-d.htm

At the time of writing (September 2022), project PIDs are rarely used in Australia. Given the ARDC's commitment to rapid development of the RAiD service, coupled with high international demand and the impending ISO certification, it is anticipated that adoption rates will increase over the coming years.

4.6. Instruments

Persistent identification and discovery of research instruments is a topical issue in Australia. The Research Data Alliance PIDINST working group[14] is developing a community-driven solution for the unique identification of research instruments. Recently, the group published a metadata schema,[15] which is being implemented by DataCite[16], to enable DOIs to be assigned to instruments. ARDC facilitates an Australasian Community of Practice for individuals interested in or working on solutions for instruments in the PID space[19]. This Group has developed use cases and best practice guides that they contribute to the PIDINST working group, however, there is no reliable data on how many research instruments there are in Australia nor is it yet a common practice to assign PIDs to these.

4.7. Samples

There is growing interest in PIDs for research samples and resources[17]. The IGSN[18], was developed in 2007 to identify individual physical samples and make them discoverable and accessible, primarily in the earth sciences. Given Australia's rich and diverse geology and ecosystems, and its strength in ecology and the geo-sciences, IGSNs are of particular interest for researchers (see section 8).

According to the GeoForschungsZentrum (GFZ) data service[20], approximately 5.3M IGSNs have been registered through three Australian registration agencies:

ARDC	3201
CSIRO	32694
Geoscience Australia	5,269,509

Table 1: The number of IGSNs that have been registered through each of three Australian registration agencies.

It is impossible to know how many physical research samples exist within the Australian ecosystem, as such estimating the percentage of adoption is impossible. However, it is clear that IGSN is important to Australia's research strategy goals and should be supported.

[19] More details can be found here: https://sites.google.com/ardc.edu.au/i4ioz/home
[20] The GFZ data service for IGSN is available here: https://dataservices.gfz-potsdam.de/igsnstats/

4.8. Implications

Where we were unable to find sufficient data to quantify an entity (such as scientific instruments), we did not include it in the cost-benefit analysis. Improving PID coverage of these entities will have the benefit of adding reliable data about the scale, nature, and extent of the return on investment that they enable as well as contributing to research integrity and reproducibility. Section 8 is a case study based on qualitative interviews about the use of PIDs for data and linked data at TERN, as well as plans to implement IGSN identifiers for geological samples, and this study shows some of the potential value for these PIDs for critical use cases.

Where there are major gaps in coverage, the process of improving the completeness of PID registries will also generate more robust estimates of the likely total number of such entities in the Australian research ecosystem and support better analyses of the reach and impact of investment in these entities.

Based on the scale and coverage assessments outlined in this section, the decision was taken to focus on three entities for the cost-benefit analysis in section 6, these are grants, projects, and publications. These were viewed as the most amenable to the benefit quantification approach used, which assesses the time taken to manually input data about an entity in research-performing organisations. However, it is important to note that there are also many benefits from interactions between PIDs (as in section 7) and these may in some cases act as an additional multiplier to the benefits of certain PIDs. The time savings achievable by comprehensive linking of grant DOIs to publications or instruments is a good example of this. Finally, while the analysis does not quantify the benefits of ORCIDs or Research Organisation Registry IDs (RORs) for researchers and institutions respectively, good coverage of these PIDs is a precondition for many of the benefits we quantify to be realised.

5. Levels of metadata reuse

With assistance from the Australasian Research Management Society (ARMS) and the Council of Australian University Librarians (CAUL), we distributed an online questionnaire to research-supporting staff at Australian universities. It included questions about the number of processes that metadata about grants and publications is used in, the number of times metadata about these entities is manually entered into a system, and by whom.

We received 27 responses from 21 unique organisations; five institutions responded twice, so their responses were merged and deduplicated for the analysis, and one response was excluded because it was incomplete.

5.1. Reuse of grant metadata

Of the 23 responses relating to grants, 19 were complete. Of these, 18 reported that information is recorded centrally by the institution when one of their researchers is awarded a grant, and one respondent indicated that this data is collected in some cases.

Across these 19 institutions, metadata about grants was accessed in an average of 5.9 separate processes (Mode: 6, Median: 6); and of the 16 institutions for which a response was completed, metadata about grants was rekeyed an average of 3.25 times (Mode: 5, Median: 3.5).

Respondents from 12 institutions reported that researchers enter grant data into systems manually, with data added 1.8 times on average (Mode: 2, Median: 2). Likewise, for the 13 institutions reporting that administrators are responsible for entering grant data, the average was also 1.8 (Mode: 2, Median: 2).

5.2. Reuse of publications metadata

Of 23 responses relating to publications, 20 were complete. Of these, 19 reported that information is recorded centrally by the institution when one of their researchers publishes a research output, like an article, review, or book. One response indicated that such data is collected in some cases.

Of the 21 institutions for which a response was completed, metadata about publications was accessed in an average of 7.3 separate processes (Mode: 8, Median: 8). In addition to a pre-populated list of activity types, several types of data use were reported under the category of 'other', and these were often reported by more than one institution. It is therefore likely that these additional activities are undertaken at many, if not most institutions, so these numbers are probably an underestimate.

Of the 19 institutions for which a response was completed, metadata about publications was rekeyed an average of 3.1 times (Mode: 2, Median: 3) by either researchers (1.6 times on average: Mode: 2, Median: 2) or administrators (1.5 times on average: Mode: 2, Median: 2).

5.3. Observations on survey responses

The respondents likely underestimated the number of reuse events and processes requiring metadata, based on the observation that respondents from the same universities often gave different answers. This suggests that no one person has visibility over all events and processes. As noted, several of the 'other' types of data use are mentioned at one or two institutions, but are likely to be undertaken at many, if not most, on top of what they reported. For the five institutions that returned two responses, the processes listed in each response did not match. In each case, the second response added additional processes. If

this phenomenon was repeated across a majority of institutions, then the average numbers would be significantly higher.

It is likely that this mismatch in responses is due to the different views on processes and priorities from the library and the research office. This indicates the importance of aggregating data from multiple departments across an institution to generate further numerical analysis and subsequent priorities for any PID strategy or implementation work.

Free text responses mention 'updates' and 'edits' to metadata. We did not attempt to quantify these, but they are an additional time and effort burden over and above what is reported here.

6. Time and financial benefits of PID integrations

Estimating the benefits of investment in scholarly infrastructure and PIDs is non-trivial. PIDs offer a combination of benefits relating to a number of areas. Their persistence protects against digital obscurity through link-rot[19]. The use of open metadata about objects and the relationships between them[20] improved research integrity and enables better research assessment and strategic decision-making[7]. The standardisation of metadata schema and central storage in registries, combined with the use of Application Program Interfaces (APIs), enable interoperability between funding, research management[21], and publishing systems, reducing administrative burden. In this section, we focus on the latter, showing how investment in research infrastructure and PIDs will result in significant time and cost savings for the Australian academic research sector.

Most of the immediate benefits of investing in PIDs lie in the simplification of research management workflows and a reduction in administrative burden, particularly for time-poor researchers[22][23]. For instance, a researcher applying for funding via the ARC application system (RMS) will benefit from its integration with ORCID (see section 7). After giving appropriate permissions, the researcher's publication and employment records can be automatically transferred from their ORCID record to their RMS profile, saving them the time, effort, and frustration of documenting their career in yet another system. If the researcher has also set up ORCID's auto-update integrations with PID services like Crossref and DataCite, their research outputs will be automatically added to their ORCID record, and then their RMS profile, further reducing both their administrative burden and the risk of errors as a result of rekeying information.

Opportunities for other time- and cost-saving integrations exist throughout the research management lifecycle and in research communications, including populating institutional Research Information Management systems (RIMS); reviews for hiring; promotion and tenure assessments; information-gathering for reporting exercises like the ERA; progress reports on individual grants for funders; and the population and maintenance of project and staff profile pages. In this analysis, we quantify the potential benefits of metadata transfer automation for three types of entities: publications, grants, and projects. Many other types of entities can have PIDs associated with them, including institutions, funders, instruments, samples, supplies and patents, however, these three entities have the most complete data available (see section 4.

We relied on previous work to estimate the time taken for manual data entry, and for the number of data entry events for project metadata. Work by Research Consulting in the UK measured the time taken to input metadata for journal articles[24], which they found to average 6.73 minutes of staff time per metadata record. In addition, we assumed that for publications, data would have to be rekeyed for each author with an average number of authors per publication of approximately four, based on the work of Fanellu and Larvière[25].

A study undertaken in Norway found that it takes an average of 10 minutes to input basic descriptive information about a research project, and that this data is typically entered into six separate systems[26]. For grants, we assumed that the time burden associated with entering information is similar to that for projects, as they both represent multi-faceted objects with the potential to encompass multiple outputs, people, and institutions, so we used the same time estimate of 10 minutes for data entry relating to funding awards.

Salary costs for three typical roles (a senior researcher or Principal Investigator, a junior researcher, and a research manager) were obtained by averaging data from NHMRC support funding models; job adverts for relevant roles (taken from www.indeed.com and www.au.talent.com); and exemplar universities of varying sizes and intensity of research activity (University of Queensland, Macquarie University, and Edith Cowan University).

Our savings calculations used an algorithm in which the number of entities of any given type are multiplied by the number of data entry events, multiplied by the time taken in minutes for data entry. This gives us an estimate of the time savings automated metadata reuse using the records associated with PIDs could bring. These aggregated savings in minutes were converted to days, assuming 7.25 working hours per day. We then multiplied the time savings by average salaries to derive financial savings.

Costs for centrally provided services were calculated using budget data shared by ARDC for the full suite of PID services they offer, including Digital Object Identifiers (DOIs) via DataCite, Handles, International Generic Sample Numbers (IGSN), Research Activity identifiers (RAiD), and Persistent Uniform Resource Locators (PURLs); and by AAF for the Australian ORCID consortium, which includes ORCID membership and support services cost recovery.

6.1. Total possible direct savings

Table 2 shows the total amount of both the time and financial cost of manual entry of metadata associated with publications, grants, and projects. The figures here represent the upper-bound of direct savings possible if all 42 universities in Australia fully integrate all priority PIDs. As can be seen in the table, the total time cost of rekeying information about publications, projects and grants amounts to 28k person days, the equivalent of nearly $24M.

Potential annual total sector savings								
	Number	*# authors*	*# rekey events*	*# minutes / event*	*cost / author / minute*		*Time savings per year (person days)*	*Financial savings per year*
Publication metadata	180,000	4	3.1	6.73	$1.13	- if Admin	**34,532**	$17,037,747
					$1.06	- if Junior Researcher		$15,972,888
					$2.13	- if Senior Researcher		$31,945,775
						Average		**$21,652,137**
Grant metadata	6,000	-	3.25	10	$1.13	- if Admin	**448**	$221,176
					$2.13	- if Senior Researcher		$414,705
						Average		**$317,940**
Project descriptions	21,083	-	6	10	$1.13	- if Admin	**2,908**	$1,434,817
					$1.06	- if Junior Researcher		$1,345,141
					$2.13	- if Senior Researcher		$2,690,282
						Average		**$1,823,413**
Total predicted annual savings from auto-feed of key metadata via API links:							**37,888**	**$23,793,490**

Table 2: The results of our analysis, showing the total annual cost in terms of both time (nearly 38k person days) and money (nearly $24M) of manually rekeying metadata about publications, grants and projects by researchers and administrators at Australian universities

6.2. Partial benefits and network effects

The level of system-wide benefit will not be linearly proportional to the adoption percentage, because the more integrations that occur, the more data is available and, by extension, the more valuable future integrations become[27]. As adoption reaches high levels, the increase in benefit will begin to flatten out, as new integrations will add fewer new PIDs and associated metadata. As this happens, the new integrations continue to benefit the individual institution, but will have less cumulative effect on the entire network. We have modelled this behaviour using a logistic function (Figure 4), which is commonly used to model such network effects due to technology diffusion[28].

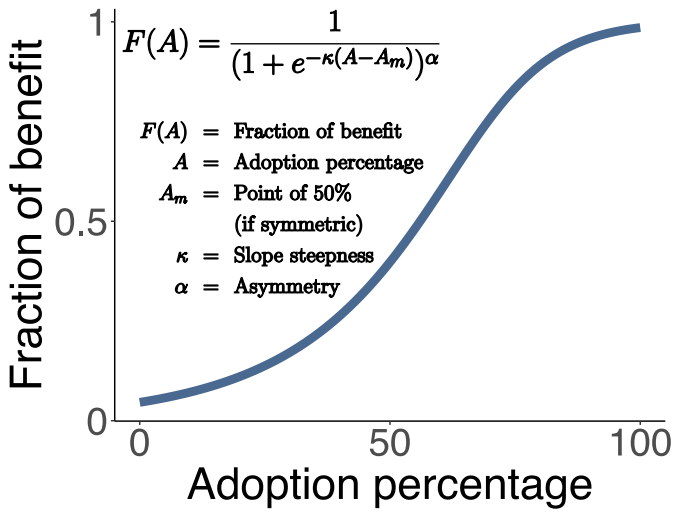

Figure 4: The individual benefits of adoption for each organisation will be greater as more organisations become part of the 'network'

Savings levels based on levels of adoptions						
Institutional adoption levels	**0%**	**20%**	**40%**	**60%**	**80%**	**100%**
Realised benefit	0.0%	1.8%	11.9%	50.0%	88.1%	100.0%
Effective time savings (person days)	0	681	4,516	18,944	33,372	37,888
Effective financial savings ($ millions)	$0.00	$0.43	$2.84	$11.90	$20.96	$23.79

Table 3: The level of benefit that is accrued in terms of both staff time (person days) and financial cost ($ millions)

Table 1 shows how financial and time savings do not increase linearly across the academic system as more universities and institutions adopt and integrate PIDs because, for maximum benefit to each university or research institution, there needs to be widespread adoption of PID workflows.

In practical terms this means that, for the greatest benefit, a high proportion of universities and other institutions need to invest in integrations and adopt PIDs as part of their workflows—PID adoption is a collective action problem requiring both community organisation and collective investment.

6.3. Costs of implementation

The potential cost savings possible from adopting the five priority PIDs are significant, but no change comes without a cost of its own. In 2015, Jisc in the UK conducted a cost-benefit analysis for an institutional implementation of ORCID. They found that it takes an average of 40 person days for a full implementation[29]. It would be tempting to simply multiply that figure by the number of PIDs covered by the national strategy and use that as the total cost of implementation but, in reality, implementation costs would be lower due to a number of factors.

The original Jisc ORCID CBA included several line items that are less relevant today, in the Australian context, than they were in 2015, in the UK. For example, the creation of promotional material to educate researchers on ORCID would not be needed today, nor would there need to be more training than at present to educate researchers on research management workflows and reporting requirements. Some of that training may need to change to adapt to new workflows but, in an PID-enabled workflow, the burden would be much lower, likely with a resulting reduction in the training needed.

The most important factor limiting costs comes down to the mode of integration. In Australia today, information systems like RIMS and repositories have been almost ubiquitously adopted[30] — products that often carry six- or seven-digit price tags. The five priority PIDs would need to be integrated at the wholesale level, through API integrations with those RIMS providers. As an example, at the time of writing (September 2022) the integration of RAiD into the Research Data Box (ReDBox) system (used by 10 Australian universities), to enable creation of RAiDs directly in the ReDBoX environment, is estimated to cost around $30, 000, which is born by the system providers. Once integrated, that functionality will be available to all institutions that use the ReDBoX product.

A vital aspect of an Australian PID strategy must therefore be to work with providers of RIMS that are popular in Australia, like Symplectic Elements, Elsevier Pure, and ResearchMaster, as well as the communities that support open source offerings like ReDBox, and VIVO, to encourage wholesale adoption of priority PID services.

6.4. Implications of analysis

Given the non-linear increases in benefits across institutions, and the complexity of subsequent network effects, three important questions to consider when designing a national Australian PID strategy are:

1. Of the 38k person days and $24M dollars currently being wasted, what target levels of adoption are required to realise enough of those potential savings?

2. What are the indirect benefits, or opportunity costs, associated with the loss of productivity caused by valuable researcher time being spent entering metadata into systems?
3. What are the costs associated with implementing these PIDs at universities, and what are the costs and benefits associated with centralised services from organisations like ARDC and AAF?

If the PID strategy can offer a compelling answer to these questions, it will be extremely helpful in motivating the sector to address the collective action problem of early PID adoption.

7. Case Study 1: ORCID and Crossref integration in ARC's RMS system for grant applications

This case study examines the dramatic reduction in tedious collection, curation, and typesetting of a researcher's publication history that has been achieved through enhancements to the Research Management System (RMS) system launched by ARC in 2018, through the lens of Joe Shapter's personal experience of how much easier it became to submit research grants and show evidence of widespread uptake of the workflow.

In 2014, ARC conducted a survey of Deputy Vice Chancellors of Research at Australian universities, which found that the ORCID was the most commonly adopted and preferred method for identifying people. These findings, coupled with a sense of momentum around ORCIDs and PIDs in general, and a government focus on the reduction of red tape[31], led to ARC's participation as a founding member of the Australian ORCID consortium, which launched in early 2016.

The vision behind driving the ARC's adoption of ORCID was to increase efficiency, reduce duplication of effort, improve data quality, and create consistency in presentation and collection of data to facilitate synthesis and analysis. As Prof. Shapter was quoted in the 2021 interview with AAF[32], applying for a grant from ARC would take 'a few weeks'. To understand why, it's important to note that researchers don't only have to maintain a list of publications, grants, projects and collaborators, but also a record of how each of these entities relate to each other. As it was put in the interview

> *Even if your CV was already up-to-date there "was still a pile of work to do" for example, "you might have to say what past grant funded the work or which of these papers that you're listing is related to the current application"*

7.1. Improved Research Management System (RMS) workflow

ARC launched a full ORCID integration within RMS in 2018, with a focus on streamlining the system to reduce administrative burden, particularly for researchers. Information about a researcher's publications can be imported directly from their ORCID record once they have verified their ID and given permission for the data to be shared. Once the connection between the RMS and the ORCID ID has been established, any future publications added to the researcher's ORCID record will automatically populate their RMS profile. A representation of the workflow is shown in Figure 5

This functionality is enhanced by ORCID's auto-update system, which has, to date, enabled approximately 5M articles globally to automatically populate ORCID records. When a

publisher, repository, or preprint server registers a DOI with DataCite or Crossref[33][21], the associated metadata flows from the publisher's systems through Crossref, into ORCID, and finally to RMS. This reduces the need for manual intervention, while still providing researchers with the ability to control when data is synchronised to their RMS account.

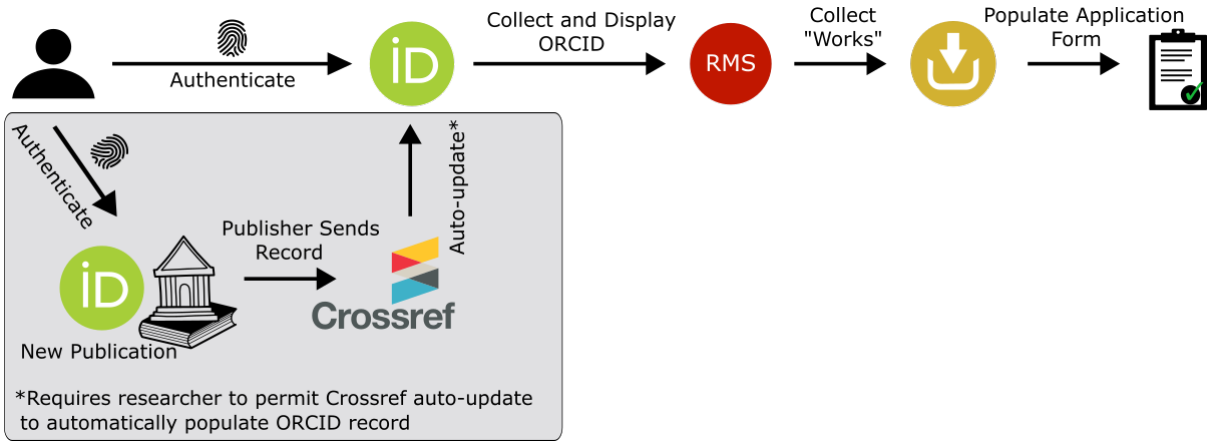

Figure 5: Researchers can import their ORCID publication records directly into the RMS system that was launched in 2018. Coupled with Crossref auto-update, metadata can travel from the publisher's systems, directly to ORCID, and then to the RMS system with no manual keying

The RMS researcher profile can, in turn, be used to automatically populate research grant applications. This workflow ensures that data only has to be imported once and can be repeatedly reused—in stark contrast to the pre-2018 system, which required researchers to compile lists of publications and submit them separately in PDF format for each application.

In addition to the ORCID import workflow, publication records can be imported as BibTeX files, or via DOIs, with the system harvesting metadata from Crossref as well as manually. While not as automated as the ORCID integration, this workflow has been especially useful for final project reports.

7.2. Driving adoption of ORCID

The time saving benefits of linking their ORCID record to their RMS profile and taking advantage of the automation that provides are clear. As Professor Shapter put it in the interview with AAF, the time savings when applying for a grant from ARC are of the order of days of work that can now be spent much more productively:

> *If I want to put in an ARC grant now and include all of my research track record, it's sitting there and ready to reuse and is being continually updated. This saved me 3-4 days per grant application - the difference in workload was staggering!"*

[21] Auto-updates: time-saving and trust-building, Blog post (2020) https://support.orcid.org/hc/en-us/articles/360006896394-Auto-updates-time-saving-and-trust-building

However, those benefits can only be realised once the researcher has taken that initial step of linking the accounts. As a demonstration of how popular this workflow has become, Figure 6 shows the number of researchers that have linked their ORCID records to their RMS profiles since the launch of the system in 2018.

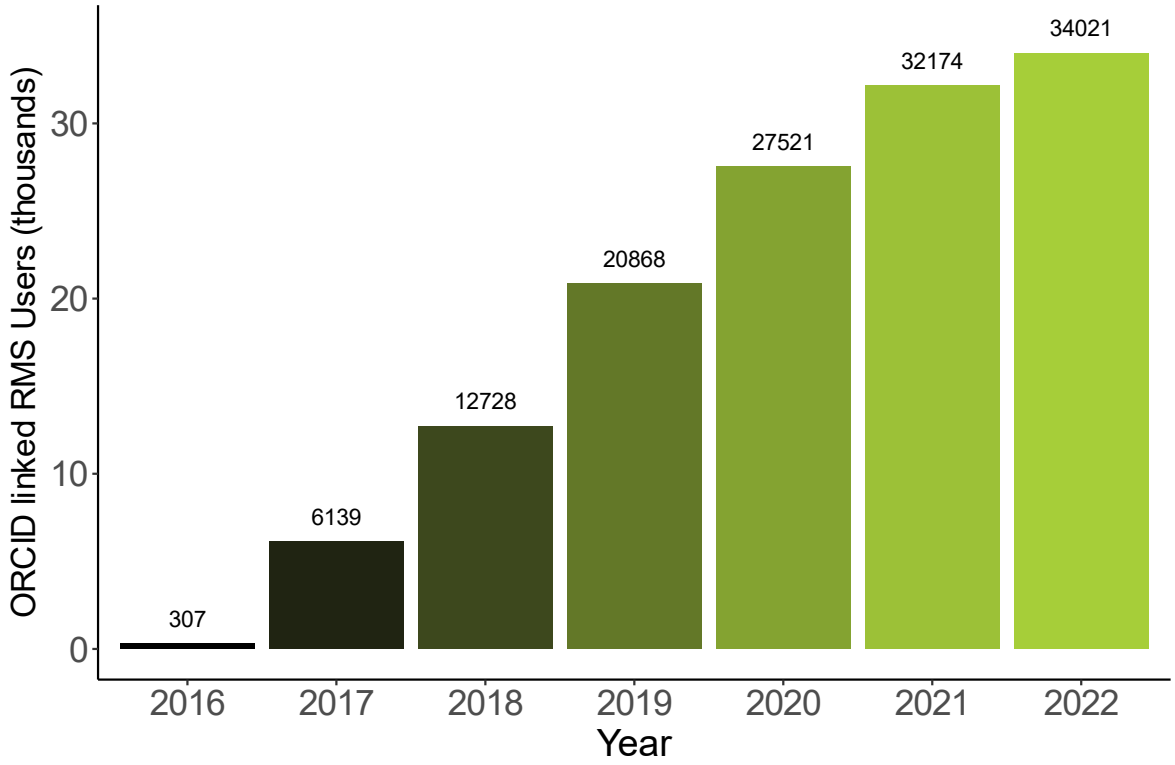

Figure 6: The number of RMS users who have linked their ORCID records.

Comparing the current number of ORCID-linked RMS users with the total number of researchers according to HERDC (108,873) suggests that as many as 31% of Australian researchers have now linked their ORCID accounts to ARC's RMS system and are using the update functionality to significantly streamline their workflow for grant applications, saving themselves time and the Australian research sector a significant amount of money.

7.3. Estimated time and money savings

According to ARC, since the launch of the new RMS system, nearly 2M publications have been imported via the ORCID API. Figure 7 shows the breakdown of the types of outputs that have been imported. The vast majority are peer-reviewed journal articles with a mix of other types of output, such as books and proceedings.

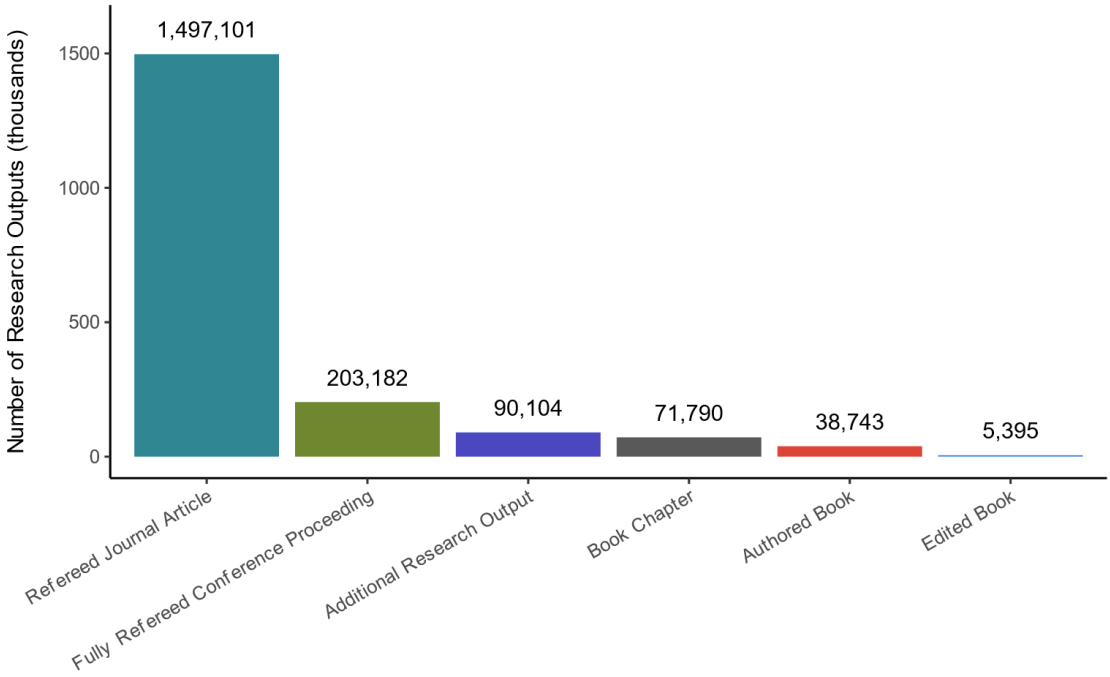

Figure 7 The total number of research outputs that have been automatically imported into ARCs RMS system. Once imported they can automatically be added to grant applications as evidence of previous work

Unfortunately, it's not possible to know what proportion of those publication records go on to be used in a competitive grant application. Based on data shown on ARCs website, 5,976 national competitive grants have been funded, out of 30,088 applications (a funding rate of 19.9%) since the launch of the RMS[22]. It is likely that not all researchers will have realised the equivalent time savings as Joe Shapter through his use of ORCID (3-4 days per application)[32]. As a senior researcher, Prof. Shapter will have a greater number of publications and previous funding awards to account for than most, but even a more modest estimate of savings of a single day of work per application suggests that, since launch, the system may have saved researchers across the sector 30,088 days of their pre-award workflows.

Additionally, at the end of the grant life cycle researchers are able to use PIDs in RMS to assist them with final reporting. Since the enhancements to RMS, there have been 4,138 final project reports filed. They included 76,145 outputs, of which 59,301 (≈78%) were submitted through the Crossref integration. This workflow simply requires researchers to enter the DOI of the publication, rather than retype the metadata. Based on our cost calculations of $14.31 per publication entered, the Crossref integration into the final report workflow has saved the Australian academic sector almost $849k and 904 days of

[22] At the time of writing, data displayed in ARC's Microsoft Power BI Tool shows a success rate of 19.9% for competitive grants with start dates form 2018 onwards: https://www.arc.gov.au/funding-research/funding-outcome/grants-dataset/trend-visualisation/ncgp-trends-success-rates (accessed Sep. 30, 2022).

researchers' time[23] that can now be spent on productive research, in addition to the savings created for the pre-submission workflows described above.

7.4. Real world evidence for time savings

It is not possible to know precisely from the RMS website data how long a researcher spends creating and submitting a grant application or, as noted above, exactly how many publications have been added to applications via the ORCID integration. Session times are not useful indicators of time and effort because they may log into the site, begin a submission, and then perhaps need to shift their attention to something else, or leave their desk for a while. However, usage data validates the assertion that researchers are now spending less time entering data into ARC's systems when applying for a grant, and that the more a researcher uses the RMS system, the more efficient the process becomes.

The median number of days between researchers beginning a grant application and submitting it is shown in Figure 8. Along the X-axis, the 'Application' number refers to the ordinal sequence of each application for each researcher, i.e., the first grant application that a researcher submits is categorised as 'Application 1', the second would be 'Application 2', and so on. Aggregating across the first six applications for each researcher, the number of days it takes to complete the submission clearly reduces over each consecutive application.

[23] Based on 6.63 minutes per publication, with the cost of a senior researcher's time being $2.13 per minute. Multiplied by 59,301 publications, that's $848,597. Time savings were calculated by multiplying 6.63 minutes by the total number of publications, assuming 7.25 hours in a working day.

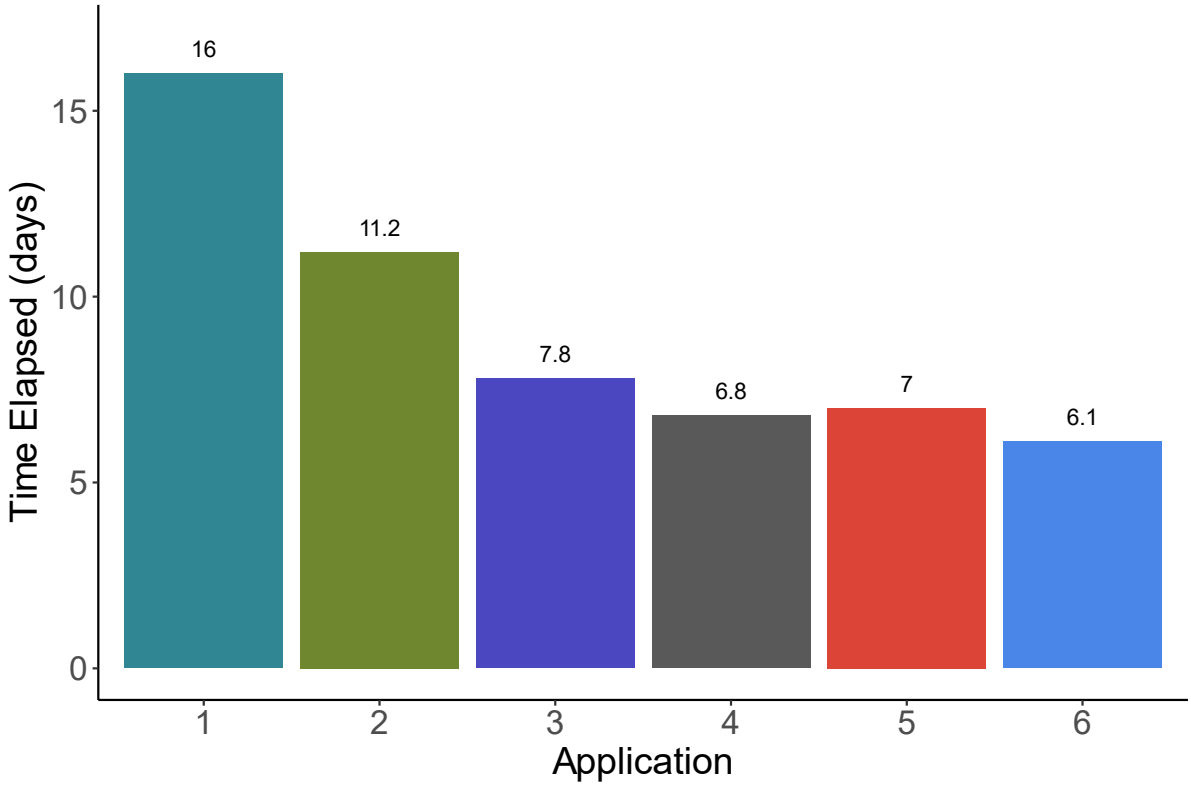

Figure 8: The median time (in days) between researchers beginning a grant application and submitting it on ARC's RMS systems. The 'Application' number '1' refers to the first time the researcher has submitted an application, '2' refers to the second time, etc.

There is a similar pattern in the number of times a researcher 'saves' their progress (Figure 9). Again, this is not a direct measure of researcher time and effort, but the reduction in the number of saves after each consecutive use of the RMS indicates that researchers may be taking fewer breaks during the process and engaging in less burdensome work.

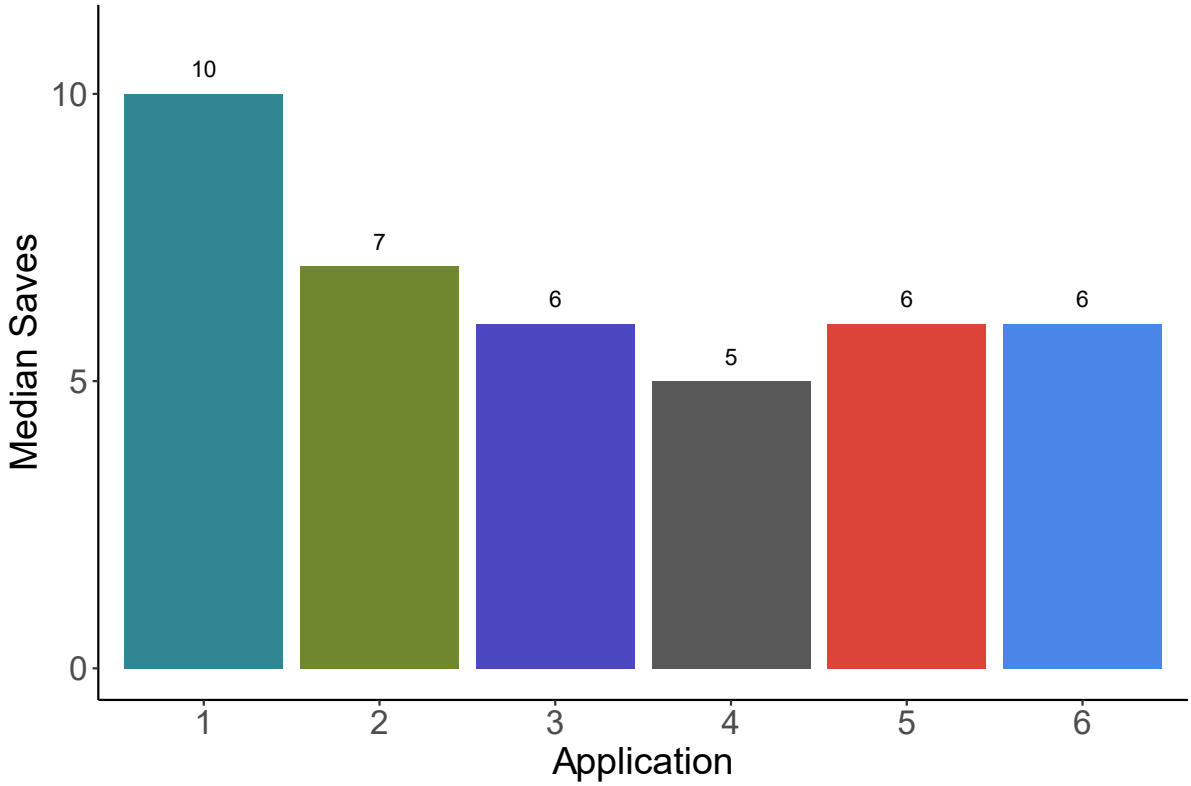

Figure 9: The median number of times that researchers saved their progress while filling in a grant application on ARC's RMS systems. The application number along the X axis refers to the ordinal sequence of the application for each researcher, that is '1' refers to the first time the researcher has submitted an application, '2' refers the second time, etc.

The reasons for longer time elapsed in completing applications and more frequent 'saves' in completing submissions are likely to be similar. They could include running out time before moving to a different activity; not having all the necessary information available; having to prepare or update documents before submission; or taking a break from the task at hand. In all cases, after an initial 'peak' in effort to get set up on the system, automatically populated data fields and extensive data re-use are clearly facilitating a more efficient application process.

7.5. Conclusion

This should not be the end of the story for ARC's PID integrations, or for funders more broadly. It is important to build on the success of this one integration to bring more automation and more time savings. As Professor Shapter notes, there is a need to link funding to publications, which requires integration with a PID for grants. In addition, PIDs for institutions and projects could streamline the project further.

This single example illustrates the impact that funder integrations can have. A national PID strategy should embrace this vision, extending it to all Australian funders as well as to

processes like the ERA and EI. Imagine the reduction in administrative burden if the ERA could be made as automated as importing a publication record into the RMS profile system.

8. Case Study 2: The use of PIDs at TERN

This case study is a qualitative analysis of semi-structured interviews with three staff members at TERN about how PIDs are currently used at TERN, how they plan to use them in the future, and the potential value that they foresee with greater PID adoption.

The Terrestrial Ecosystem Research Network (TERN) is Australia's terrestrial ecosystem observatory. Founded in 2009, it enables coordination and collaboration between the Australian national, territorial, and state administrations, over 20 universities, a variety of non-profit and commercial organisations, nearly two dozen National Collaborative Research Infrastructure Strategy (NCRIS) projects, and a number of international collaborators[24]. TERN is funded through NCRIS, as part of the National Research Infrastructure (NRI)[34], which provides site-based research equipment and delivers a series of multi-scale open data, research and management tools, as well as data infrastructure. Since inception in 2009, TERN's infrastructure has been instrumental in enabling researchers to publish over 1,000 academic articles[25].

The ecological field observatory supported by TERN is structured around three key scales:

1. Landscapes — relies heavily on remote sensing and model-derived dataset at the regional and continental scales to provide information about landscapes, topology, vegetation composition, fire dynamics, soil attributes, and many other measurements.
2. Ecosystem surveillance — plot-based field measurements and samples provide information about the direction and magnitude of changes to Australia's environment. Core attributes observed are from vegetation and soil domain
3. Ecosystem processes — long-term, time series data for detecting changes to Australia's delicate ecosystems. Observations are conducted at key representative ecosystem sites such as Alice Mulga at Pine Hill Cattle Station in the Northern Territory, and the Warra Tall Eucalypt, a world heritage site in Tasmania that is traditionally owned by the Palawa and Pakana people

8.1. Shared resources to support collaboration

The research conducted through TERN is necessarily highly collaborative and extremely diverse. To facilitate this collaboration, TERN maintains a data portal called Submission, Harmonisation, and Retrieval of Ecological Data (SHaRED) supported by a dedicated infrastructure and software team. The TERN data portal contains time-series remote sensing and in-situ sensors data sets that are released at regular intervals, as well as static, one-off

[24] A list of TERN's partners can be found on the website here: https://www.tern.org.au/partners/
[25] A list of publications enabled by TERN's infrastructure contain nearly 2,000 records at the time of writing (September 2022) https://www.tern.org.au/research-publications/

data sets and a range of other entities such as vocabularies, traits, instrument records, methods, organisations, and people. Data types include traditional tabulated data, images, acoustic files, geospatial vector and raster data, and several others.

In addition to the SHaRED data portal, a trusted research environment via a cloud-based virtual desktop, enabling researchers to remotely connect to powerful computational resources from anywhere in the world.

With such a complex, collaborative research ecosystem, keeping track of people, places, and things can prove administratively challenging and time-consuming, with a real risk of errors and miscommunications. As a result, TERN employs data scientists, data librarians, and engineers to support researchers in the use of technology, including PIDs, to enable more efficient collaboration.

8.2. Research data sharing is in evolution

The researchers we spoke to at TERN say that they have seen a dramatic change in research data sharing over the last decade. Plant ecologist Dr Lachlan Charles said:

> *I think there's … push from journals. When you're publishing a paper it seems to be more apparent in the last 10 years. [Publishers ask you] to upload your data on Dryad or one of those kinds of repositories… They will mint a DOI for you as well. So there seems to be that push at least in [the] plant community ecology, it's quite common that it's actually a requirement now.*

However, not all fields have made the same level of progress towards open data. Dr Elisa Maria Girola, a bio-acoustics expert, told us about her experience of being unable to find data because of a lack of PIDs:

> *In the field I work in (bioacoustics)[...] it's not yet at the point where [registering DOIs for data is] common practice. … In the past, I've struggled to find data that were supposed to be there and were not. I think having more DOIs will overcome this issue because if it has a DOI, it's usually still accessible.*

When asked about the benefits of DOIs for data, Dr Girola said that she was really looking forward to greater adoption in order to see substantial benefits, like the ability to track data set reuse in the same way that hypotheses and ideas can be tracked by bibliometricians using citation analysis today:

> *I think it is something that we're going to see further down the line in a few years time… when the practice of using persistent identifiers specifically to identify data sets is a bit more commonly used and we start to see the same data set being*

> *reused in different work. Then you can start tracking the use of that particular data set across different publications or different projects.*

As well as avoiding loss of data that is no longer stored at its previous or cited URL, PIDs provide a more direct and more robust way of finding data sets. Dr Charles talked about the circuitous, and often error-prone, approaches that researchers are forced to take in the absence of DOIs to try to find research outputs, such as painstakingly trawling through repositories, or using services that are not designed for the purpose, like general open web search engines (e.g., Google).

> *It's just faceted search on whatever portal or website [you think the data is on. You] type the keywords and off you go. [Often] you have to … copy [and] paste the title [of the dataset into a search engine] but if you [have] the DOI, you go straight to the source.*

TERN uses the ARDC-hosted DOI service to register DOIs for static datasets and data products. Users can request a DOI when submitting their data through SHaRED. TERN is recognised as one of the preferred repositories for hosting ecology data for Nature Scientific data[26].

8.3. Linked data and interoperability

TERN makes use of the W3ID[27] permanent identifier system to ensure stable URLs for linked-data resources like vocabularies, taxonomies, and ontologies. W3IDs provide direct deep links to data sets where the URL will not change, even if the internal structure of the repository is changed. That is, W3ID solves the problem caused by the inherent ephemerality of URLs (known as 'link rot'), with metadata associated with the dataset provided by the repository itself.

Dr Charles summed up how persistent identification of datasets makes research more efficient:

> *For my job, it helps [to have] persistent identifiers, because you [can] track back to the source [of the] data. For example, if we have a terminology that we've taken exactly off an ontology website, we link that to that site and to that persistent identifier…. So it helps just in following the breadcrumbs to … where it came from.*

In much of the work supported by TERN, ontologies and taxonomies are important. In the fields of ecology, plant science, climate science, and geology, objects need to be catalogued,

[26] https://www.nature.com/sdata/policies/repositories#ecology

[27] More information about W3IDs, their purpose, management and links to relevant organisations can be found here: https://w3id.org/

classified, and defined so that it is clear to everyone exactly what is being referred to. Dr Girola explained how, particularly for newer disciplines, PIDs are vital to avoid confusion:

> *A lot of fields, especially in ecology and biology, are rather new and definitions can be confusing or misinterpreted. They've not been around long enough to be set in stone. …Persistent identifiers [that] have the definition and the source allows us to say that [a] parameter that is in our data set has a certain name.*

In addition, the use of PIDs enables interoperability for data sets that include definitions and taxonomies, a key component of the FAIR[1] principles for data management. As Dr Girola put it:

> *[For new research we can] use the same label and refer to the same persistent identifier for the definition and source. That way everybody knows that you're talking about and it's comparable to what already exists.*

8.4. IGSN at TERN

Physical samples are critically important in fields like ecology, plant biology, geology, and archaeology. Given the collaborative nature of the work supported by TERN, it is vital to carefully curate and keep track of the physical samples that are used, shared, analysed, and re-analysed. Dr Anusurya Devaraju, Senior Data Innovation Manager, outlined the reasons why TERN is investing in IGSN:

> *Physical samples are essential research assets of TERN. Amongst other things, TERN has collected more than 100,000 soil samples, soil metagenomic samples, leaf tissue and plant specimens from survey plots across Australia. These samples are freely available to research communities, both nationally and internationally. TERN plans to implement IGSN for better management of sample collection and improve wider discoverability and accessibility. The identifiers will help the sample management team quickly manage and track samples' extensive collections and derivatives (sub-samples) and promote interoperability of sample data collection. Sample Identifiers will be linked with the datasets derived from the samples and other related datasets collected during the same site visit.*

On a practical note, Dr Charles gave an example of how IGSN would be extremely useful in a field in which he used to work: the study of invasive grass species. He mentioned how the names of new organisms sometimes change or they move within taxonomic groups as researchers learn more about them:

> *In my own work in California there's a lot of invasive grass species. …some of these were identified in the 60s and given specific names. Over time, taxonomists have … either made them subspecies or [it's been found that a] plant species has been*

called by two different scientific names. … Just searching those using the scientific name as a keyword can be problematic, because then you can [miss] half of your potential searches…. If you had specific identifiers for samples, then you can compare them in the world of taxonomy, then you could solve some problems.

8.5. Conclusion

As a vital piece of ecological research infrastructure in Australia, TERN has invested significant time and effort in building user-friendly, efficient services that enable the Australian ecological research community to collaborate more effectively and impactfully to protect Australia's fragile ecosystems.

The use of PIDs has been extremely helpful in creating permanent and unique identifiers for data sets and linked data resources like taxonomies and ontologies. As a result, the researchers we spoke to feel that PIDs enable them to find the resources they need more quickly and easily. In addition, their workflows are now more robust—PIDs have reduced the risks of confusion, errors, and missed research opportunities.

There is still more work to be done. Many of the benefits that researchers see from PIDs are yet to be fully realised, as adoption is highly variable across disciplines and communities. With adequate support, the future of collaboration and interoperability will be even brighter at TERN. The adoption of IGSN promises to reduce risks of confusion and streamline collaboration with respect to physical samples.

9. Case Study 3: National-scale, centralised PID services

This case study examines benefits of central PID services for institutions, using the AAF ORCID Consortium and ARDC PID services data as exemplars.

A national PID strategy involving key stakeholders in the Australian research sector, including universities, research institutions, funders, and infrastructure providers with a strong adoption target, would save significant amounts of both time and money. These benefits are not abstract but will make a real and tangible difference to researcher's daily work and quality of life. For example, as described in section 7 and illustrated by Joe Shapter in his ORCID story[32] the adoption of ORCID into ARCs systems and workflows has already made a huge difference.

For institutions that want to adopt PID-enabled workflows in order to realise similar benefits, there are some perceived barriers to implementation, in particular, the cost of implementation.

As described in section 6.2, the case for an individual university to invest in integration becomes more compelling as the number of integrations across the system increases. There is a sort of collective action challenge at work, which can be helped by the provision of centralised services as part of a national strategy. These strategic services can lower up-front costs, reduce implementation time, develop a sense of community, foster best practice, and encourage system-wide adoption.

9.1. Services provided by ARDC

ARDC has been a world-leading advocate and provider of a range of PID services to the academic sector in Australia for over 10 years. The five main areas of service are:

- DOIs - as a DataCite registration agency, ARDC supply DOIs for research data, software, grey literature, and instruments
- RAiD - a new type of PID for research projects
- IGSN - for physical objects collected during the course of research
- PURLs - persistent URLs for research grants
- Handles - a general purpose identifier, and the foundation of RAiD

9.1.1. DOIs from Datacite

ANDS, one of the three organisations that came together in 2018 to form ARDC, was a founding member of DataCite. ARDC has continued that legacy of pioneering work

including, in 2014, creating a web interface for DOI registration called *Cite My Data*, a service that was a forerunner to the DataCite's Fabrica[28].

While DataCite manage the DOIs, since inception, ARDC's DataCite service has provided the infrastructure to allow registration, resolution, and updating of 610,364 DOIs for 59 publicly funded research organisations and data centres relevant to Australian research, such as those run by government agencies or universities, with a further seven organisations developing integrations at the time of writing.

9.1.2. Research activity identifier

RAiDs are a new type of PID for projects, which is being developed by ARDC[29] to provide a central minimal information set to describe research projects and a metadata "envelope" that includes a set of date stamps for the associations between the described project and other PID-described entities:

- Funders
- Organisations
- Collaborators
- Tools and services
- Data

Projects are the fundamental unit of research activity. The accurate, timely, and complete description of their associations, inputs, and outputs are critical to research management and reporting including, for example, for the ERA. Their central importance is evidenced by the rapid rise and near universal adoption of RIMS, like Symplectic Elements, Elsevier Pure, or ResearchMaster[30].

As noted in section 4.5, our estimate for the number of research projects in Australia is based on scaling an estimate of the number of UK projects by the relative levels of research investment between the two countries. This was necessary because the databases of projects held in RIMS are proprietary, and not open without a specific effort by universities to make this information publicly available.

RAiD offers an opportunity to create a strong unified standard on metadata for research projects and a central registry of that metadata, thereby simplifying its aggregation, sharing, and reporting, by making it 'portable' so that it can more easily be moved between systems, for example, when a university changes RIMS vendors—with the caveat that not all project information can be made openly available for commercial or legal reasons. There is an

[28] For more information about Fabrica, see DataCite's website. https://doi.datacite.org/
[29] The RAiD website can be found here: https://www.raid.org.au/

expression often used in open research to address this concern; project information should be *'as open as possible - as closed as necessary'*.

RAiD has emerged from a series of pilot implementations at universities in Australia and is currently being formalised as an ISO standard. A global network of partners and advisors, including likely users, potential registration agencies, and domain experts, is currently feeding into the next generation of the metadata schema, use cases, and sustainability models.

9.1.3. International Generic Sample Number

ARDC is one of four IGSN allocation agents in Australia (alongside Geoscience Australia, Commonwealth Scientific and Industrial Research Organisation (CSIRO), and Lithodat). IGSNs were originally developed for use by the Australian earth science research community and can be applied to physical samples, including- for example geological soil, rock, and sediment samples.

As mentioned in section 4.7, ARDC has provided over 3,000 IGSNs for a range of organisations and is currently looking to expand their use by working with research communities beyond the geosciences.

9.1.4. Persistent Uniform Resource Locator

ARDC provides PURLs as a unique identifier and persistent resource location service. To date, there have been 33,201 PURLs issued on behalf of NHMRC and 29,473 issued on behalf of ARC. Grant identification is a long-standing practice at ARDC, but currently PURLs are issued to only about one quarter of research and development grants nationally, according to ARDC's estimates. As discussed earlier, PURLs do not capture grant metadata in the way that DOIs do, therefore a transition from PURLs to DOIs, in line with current international best practice, is recommended.

While ARDC uses DOIs to identify its own co-investment with partner organisations and has registered 77 grant DOIs at the time of writing, other research funding programs such as ARC's Discovery Program are yet to adopt this practice. Given the scale of research funding identified in section 4.3, with 6,000 or more grants issued annually to Australian research-performing organisations, adoption of grant DOIs, should be considered a priority. While the analysis in section 6 focused on the time savings for entering data about grants alone, far greater savings could be achieved by linking those grants to people, outputs, facilities, and so on. Adoption of grants alongside these other PIDs could be transformative in a way that grant DOIs alone will not.

9.1.5. Handles, the general-purpose identifier

ARDC offers a machine-to-machine minting service for Handles that allows easy implementation into repository and other university systems. In addition, a batch-minting service and a self-minting interface is offered that is accessible through the user's AAF account, for those who have not yet set up the machine-to-machine integration.

Through these channels ARDC have provided 439,203 globally unique, citable identifiers to 20 organisations for datasets, collections, papers, and other research outputs.

In addition, the Handle service provides the underlying technology that underpins both the IGSN service and RAiD.

9.2. Costs of ARDC Services

According to ARDC budget figures, the operating costs of all these PID services comes to a little over three quarters of a million Australian dollars, as shown in Table 4*: A breakdown of ARDC staffing, licensing and strategic projects costs. Note that equipment and systems are not costed as they are covered by the 'on costs' portion of staffing. RAiD is a dedicated development project within ARDC and therefore costs are not included here.*

ARDC PID Services	Cost per year
ARDC - PID services staff costs	$540,000
ARDC - Licensing, consortium, membership fees	$75,000
ARDC - Strategic projects and incidentals	$150,000
Total costs for ARDC PID services	**$765,000**

Table 4: A breakdown of ARDC staffing, licensing and strategic projects costs. Note that equipment and systems are not costed as they are covered by the 'on costs' portion of staffing. RAiD is a dedicated development project within ARDC and therefore costs are not included here.

The levels of costs for these services are surprisingly modest, given the level of value currently being offered, particularly when set against the $24M of administrative burden identified as part of this report (section 6).

9.3. The national ORCID consortium, managed by AAF

In this section, we look at a specific example of the ORCID consortium service provided by AAF and show the effective cost savings that it delivers.

Forty-three Australian institutions (mainly universities, but also research organisations and funders) are now members of the AAF-managed national ORCID consortium. At the institutional level, this has led to annual savings through substantial reductions in member costs, at the same time as administrative burdens have been decreasing. Table 5 gives the annual ORCID membership cost to individual institutions from 2016 to 2022 and illustrates how the Consortium model cut costs by well over 50%, with per member savings of approximately $108k, and a total saving to all members of approximately $4.6M across this time period.

	via Consortium	Via ORCID	Savings / Member
	(AUD)	(AUD)	(AUD)
2016	$11,454.85	$27,894.00	$16,439.15
2017	$10,912.19	$26,109.66	$15,197.47
2018	$12,270.74	$26,546.39	$14,275.65
2019	$13,621.12	$29,096.05	$15,474.93
2020	$13,886.22	$29,137.20	$15,250.98
2021	$13,511.60	$29,137.20	$15,625.60
2022	$14,310.18	$30,011.32	$15,701.14
TOTAL	$89,966.90	$197,931.82	$107,964.91

Table 5 shows the difference in annual ORCID membership cost to individual institutions between the Consortium or direct membership options

9.4. Governance of the AAF ORCID consortium

The Australian ORCID Consortium launched in 2016 and is now approaching the end of its seventh year of operations. As Consortium Lead, AAF acts as a national broker between Australian institutions and the ORCID organisation, providing members with local support for integrations, access to reports, and information and advocacy materials to encourage take-up. The consortium model also ensures that the Australian research community has a unified voice within the global ORCID community. Additionally, Linda O'Brien of Griffith University has served as Chair of the international ORCID Board since 2020 and in this role she provides a direct link to ORCID for the Australian Consortium Governance Group, which she also serves as a member of.

9.5. Shared services

Adopting PID-enabled workflows (Figure 5) has been proven to save considerable time for researchers. The shared services and consortium models have already been embedded by both ARC and TERN, and return significant cost savings to institutions, and research quality benefits more generally. The logical conclusion is that a shared services model which delivers a range of PIDs to Australian organisations is likely to deliver the most cost-effective,

administratively efficient, and speedy benefits to researchers, research administrators, and directly to institutions.

10. Discussion

Australia has an advanced, well-organised, and -resourced national research ecosystem. The 2020 review of the ERA and EI by ARC[2] highlighted that Australian universities conduct *excellent research by global standards* and that there are *outstanding examples of universities translating research into economic and social benefits*. To maintain these high standards, it is important that timely, accurate, and complete information is available to research assessors and policymakers at all levels. As the complexity of research has increased, with greater levels of collaboration[35] and cross-disciplinarity[36], the issues associated with the collection, curation and reporting of this information have become ever more challenging.

PIDs offer a mechanism for standardising and storing metadata about various entities in the research landscape. The five priority PIDs for people, institutions, grants, projects, and outputs represent a suite of interoperable and linked identifiers and information for underpinning research management and assessment workflows in ways that radically reduce the time and effort required.

Our key finding that the universal adoption of these PIDs across the Australian research sector could potentially save up to 38k person days and $24M are remarkable from a purely economic point of view. But, from a personal perspective, the importance of this work is even more compelling. Joe's story[32], referenced throughout this report, is based on an interview given by Joe Shapter, Pro-Vice-Chancellor at University of Queensland, in which he describes the use of the ARC's ORCID implementation saving him three to four days of frustrating, tedious administrative work for each and every grant application he completes. Scaling that time and effort saving for every reporting or application process, for every researcher at every institution in Australia, both paints a picture of the current waste of precious researcher time and provides a compelling vision of a more efficient future.

The success of ARC's ORCID implementation, ARDCs pioneering PID services, and the AAF-led Australian ORCID consortium have created great value for the Australian research community, and these organisations are well placed to make a key contribution to development of a much-needed national PID strategy for Australia. Such a strategy would need to address government expectations of delivering high-quality research in the Australian national interest, while also reducing administrative burden to free up researcher time, streamlining workflows, and reducing institutional costs. It would provide technical and outreach support and ensure that Australia's research community continues to be fully represented on the global research infrastructure stage. This would involve investment of time and resources but, given the clear success indicators to date, a strong call to action is warranted.

Based on the analysis presented in section 6, the greatest and most predictable benefits will come from concentrating on driving high levels of adoption and leveraging the associated

network effects of the five priority PIDs, which is why we recommend setting a strong target for sector-wide adoption over the next five years. The potential benefits of comprehensive PID adoption and coverage for entering data about just three entities (publications, grants, and projects) in our analysis demonstrate the level of return PID investments can offer, but the time scale and peak levels of adoption will change the calculations of break-even and return on investment. It is important to note that optimal benefits for these three entities still depend on robust coverage of PIDs for people and organisations (as exemplified by ARC's ORCID integration in its RMS). These considerations could be addressed with a clear roadmap that establishes targets and milestones for PID adoption and coverage for a core set of prioritised PIDs.

Other PIDs, such as those for instruments, physical samples, and reagents are also important and should be addressed as part of a medium- to long-term strategy, noting that the value of such PIDs will be greatly enhanced when introduced to a research system in which their relationships to people, grants, projects, and other key entities can be readily described and recorded

Our research has shown that there is limited data on current levels of PID adoption within universities and other research-performing organisations. AAF has a clear view of levels of ORCID integrations among consortium members but, for example, there are gaps in our knowledge when it comes to organisation identifiers. Additional research is therefore needed to establish a baseline for current levels of PID integration and adoption. Without this, the strategy risks setting impossible targets and creating unrealistic expectations. In a competitive and high-performing ecosystem, this research will need to be carefully designed to enable the fullest information-sharing with the lowest organisational risk.

It is clear from this analysis that investing time and effort in developing a national strategy, focused on delivering immediate benefits from PID adoption and targeting the creation of network effects, will deliver a significant return. Over time, it will enable benefits both broad and deep, by incorporating PIDs that enhance the coverage of the national, and global, information network, which will enable new analyses and enhance the strategic management of research.

11. Bibliography

[1] M. D. Wilkinson *et al.*, 'The FAIR Guiding Principles for scientific data management and stewardship', *Sci. Data*, vol. 3, no. 1, p. 160018, Dec. 2016, doi: 10.1038/sdata.2016.18.

[2] A. R. Council, *ERA EI Review Final Report*. Australian Research Council, 2020. Accessed: Sep. 26, 2022. [Online]. Available: https://www.arc.gov.au/sites/default/files/era_ei_ac_report.pdf

[3] S. L. Schneider, '2018 Faculty Workload Survey', University of South Florida, Research Report, 2020. [Online]. Available: http://web.archive.org/web/20201209160711/https://thefdp.org/default/assets/File/Documents/FDP%20FWS%202018%20Primary%20Report.pdf

[4] J. Miller, 'Where does the time go? An academic workload case study at an Australian university', *J. High. Educ. Policy Manag.*, vol. 41, no. 6, pp. 633–645, Nov. 2019, doi: 10.1080/1360080X.2019.1635328.

[5] Parliament of Australia, 'Inquiry into Funding Australia's Research'. https://www.aph.gov.au/Parliamentary_Business/Committees/House/Employment_Education_and_Training/FundingResearch (accessed Sep. 11, 2022).

[6] C. Brown, N. Simons, Daniel Bangert, and S. Sadler, 'National PID Strategies working group', *RDA Alliance*, Aug. 05, 2021. https://www.rd-alliance.org/groups/national-pid-strategies-wg (accessed Sep. 11, 2022).

[7] Brown, Josh, Jones, Phill, Meadows, Alice, Murphy, Fiona, and Clayton, Paul, 'UK PID Consortium: Cost-Benefit Analysis', Zenodo, Jun. 2021. doi: 10.5281/ZENODO.4772627.

[8] J. Brown, 'Developing a persistent identifier roadmap for open access to UK research'. 2020.

[9] Brown, Josh, Jones, Phill, Meadows, Alice, and Murphy, Fiona, 'The PID-optimised Research Lifecycle', Jun. 2021, doi: 10.5281/ZENODO.4991733.

[10] Brown, Josh, Jones, Phill, Meadows, Alice, and Murphy, Fiona, 'PID-optimised workflows: A vision of a more efficient future', Sep. 2022, doi: 10.5281/ZENODO.7085489.

[11] Australian Government, 'Higher Education Support Act'. 2003. [Online]. Available: https://www.legislation.gov.au/Series/C2004A01234

[12] D. W. Hook, S. J. Porter, and C. Herzog, 'Dimensions: Building Context for Search and Evaluation', *Front. Res. Metr. Anal.*, vol. 3, p. 23, Aug. 2018, doi: 10.3389/frma.2018.00023.

[13] L. Stone, 'Handle Project: Competitive Evaluation of PURLs', *Competitive Evaluation of PURLs*, 22-Mar-00. http://web.mit.edu/handle/www/purl-eval.html (accessed Sep. 11, 2022).

[14] M. Stocker *et al.*, 'Persistent Identification of Instruments', *Data Sci. J.*, vol. 19, p. 18, May 2020, doi: 10.5334/dsj-2020-018.

[15] R. Krahl *et al.*, 'Metadata Schema for the Persistent Identification of Instruments', 2021, doi: 10.15497/RDA00070.

[16] M. Buys, R. Dasler, and M. Fenner, 'PIDs for instruments: a way forward', Mar. 2020, doi: 10.5438/TDK2-2G94.

[17] E. Plomp, 'Going Digital: Persistent Identifiers for Research Samples, Resources and Instruments', *Data Sci. J.*, vol. 19, p. 46, Dec. 2020, doi: 10.5334/dsj-2020-046.

[18] A. Devaraju, J. Klump, V. Tey, R. Fraser, S. Cox, and L. Wyborn, 'A Digital Repository for Physical Samples: Concepts, Solutions and Management', in *Research and Advanced Technology for Digital Libraries*, vol. 10450, J. Kamps, G. Tsakonas, Y. Manolopoulos, L. Iliadis, and I. Karydis, Eds. Cham: Springer International Publishing, 2017, pp. 74–85. doi: 10.1007/978-3-319-67008-9_7.
[19] R. Sanderson, M. Phillips, and H. Van de Sompel, 'Analyzing the Persistence of Referenced Web Resources with Memento', 2011, doi: 10.48550/ARXIV.1105.3459.
[20] H. Cousijn *et al.*, 'Connected Research: The Potential of the PID Graph', *Patterns*, vol. 2, no. 1, Jan. 2021, doi: 10.1016/j.patter.2020.100180.
[21] M. S. Mayernik and K. E. Maull, 'Assessing the uptake of persistent identifiers by research infrastructure users', *PLOS ONE*, vol. 12, no. 4, p. e0175418, Apr. 2017, doi: 10.1371/journal.pone.0175418.
[22] 'Making the most of technology to ease the burden of research administration', *ResearchMaster*, Jan. 20, 2020. https://data.com.au/researchmaster/research-administration/making-the-of-technology-to-ease-the-burden-of-research-administration/ (accessed Sep. 05, 2022).
[23] M. Munafò, 'Universities' reliance on free labour is unsustainable', *Research Professional News*, May 03, 2022. https://www.researchprofessionalnews.com/rr-news-uk-views-of-the-uk-2022-5-universities-reliance-on-free-labour-is-unsustainable/ (accessed Sep. 05, 2022).
[24] Research Consulting, 'Counting the Costs of Open Access', London Higher and SPARC Europe, Nov. 2014. Accessed: May 09, 2021. [Online]. Available: http://www.researchconsulting.co.uk/wp-content/uploads/2014/11/Research-Consulting-Counting-the-Costs-of-OA-Final.pdf
[25] D. Fanelli and V. Larivière, 'Researchers' Individual Publication Rate Has Not Increased in a Century', *PLOS ONE*, vol. 11, no. 3, p. e0149504, Mar. 2016, doi: 10.1371/journal.pone.0149504.
[26] M. H. Klausen, 'Even Minor Integrations Can Deliver Great Value – A Case Study', *Procedia Comput. Sci.*, vol. 106, pp. 153–159, 2017, doi: 10.1016/j.procs.2017.03.011.
[27] A. Dappert, A. Farquhar, R. Kotarski, and K. Hewlett, 'Connecting the Persistent Identifier Ecosystem: Building the Technical and Human Infrastructure for Open Research', *Data Sci. J.*, vol. 16, no. 0, Art. no. 0, Jun. 2017, doi: 10.5334/dsj-2017-028.
[28] D. Kucharavy and R. De Guio, 'Application of S-shaped curves', *Procedia Eng.*, vol. 9, pp. 559–572, 2011, doi: 10.1016/j.proeng.2011.03.142.
[29] R. Johnson, H. Henderson, and H. Woodward, 'Institutional ORCID Implementation and Cost-Benefit Analysis Report', Jisc-ARMA, Jul. 2015. [Online]. Available: https://doi.org/10.5281/zenodo.1445290
[30] R. Bryant *et al.*, 'Practices and Patterns in Research Information Management: Findings from a Global Survey', 2018, doi: 10.25333/BGFG-D241.
[31] K. L. Dow, 'Review of Higher Education Regulation', *Department of Education*, Nov. 19, 2020. https://www.education.gov.au/reviews-and-consultations/review-higher-education-regulation (accessed Sep. 29, 2022).
[32] AAF, 'ORCID User Stories', *Australian Access Federation*, Aug. 04, 2021. https://aaf.edu.au/orcid-user-stories/ (accessed Sep. 21, 2022).

[33] 'Auto-updates: time-saving and trust-building', *ORCID*, 2020. https://support.orcid.org/hc/en-us/articles/360006896394-Auto-updates-time-saving-and-trust-building (accessed Sep. 29, 2022).
[34] Australian Government, *2020 Research Infrastructure Investment Plan*. 2020. Accessed: Mar. 17, 2022. [Online]. Available: https://web.archive.org/web/20211129231320/https://www.dese.gov.au/2020-research-infrastructure-investment-plan/resources/2020-research-infrastructure-investment-plan
[35] J. Y. An, R. J. Marchalik, R. L. Sherrer, J. A. Baiocco, and S. Rais-Bahrami, 'Authorship growth in contemporary medical literature', *SAGE Open Med.*, vol. 8, p. 205031212091539, Jan. 2020, doi: 10.1177/2050312120915399.
[36] J. Garner, A. L. Porter, M. Borrego, E. Tran, and R. Teutonico, 'Facilitating social and natural science cross-disciplinarity: Assessing the human and social dynamics program', *Res. Eval.*, vol. 22, no. 2, pp. 134–144, Jun. 2013, doi: 10.1093/reseval/rvt001.

Appendix A: List of Acronyms

ANDS	Australian National Data Service - One of the organisations that came together to form ARDC. Established in 2008 to make research data discoverable and accessible.
API	Application Program Interface - Technology that allows computer systems to communicate. In this report, API generally refers to web APIs that enable data to move between services and products e.g. Publications in an ORCID record imported into a RIMS.
ARC	Australian Research Council - Primary non-medical research funding agency of the Australian Government.
ARMS	Australian Research Management Society - Professional association of research management professionals.
CAUDIT	Council of Australian University Directors of Information Technology - Representative leadership body for university libraries in Australia.
CAUL	Council of Australian University Librarians - Membership organisation of University Libraries in Australia.
CSIRO	Commonwealth Scientific and Industrial Research Organisation - Federally funded research organisation.
CoESRA	Collaborative Environment for Scholarly Research and Analysis - Cloud-based virtual workbench for data analysis provided by TERN.
DOI	Digital Object Identifier - PID that can identify research outputs as well as grants.
EI	Engagement and Impact assessment - National research evaluation exercise that assesses how well researchers engage with the end-users of research.
ERA	Excellence in Research for Australia - National research evaluation exercise focussing on the academic quality of research at academic institutions in Australia, developed and administered by the ARC.
FAIR	Findable, Accessible, Interoperable, and Reusable - Set of principles for scientific data management to maximise the utility of data.
GFZ	GeoForschungsZentrum - National research centre for Geosciences in Germany.
IGSN	International Generic Sample Number - PID for physical samples to make it easier to persistently and uniquely identify them.
ORCID	Open Researcher and Contributor ID - PID for people that includes a record of funding, and associated organisations. The name ORCID also refers to the membership organisation that provides and administers the PID.
PID	Persistent Identifier.
PIDINST	Persistent Identification of Instruments - Research Data Alliance group looking into Identifiers for operational scientific instruments to help preserve experimentally relevant instrument characteristics as metadata

PURL	Persistent Uniform Resource Locator - System for persistent URL redirection developed at OCLC in 1995. It does not have an associated metadata schema.
NCRIS	National Collaborative Research Infrastructure Strategy - Program funded by the Australian government to ensure researchers have access to cutting edge national research infrastructure.
NHMRC	National Health and Medical Research Council - Main statutory authority of the Australian government responsible for medical research and primary medical research funding agency.
RAiD	Research Activity Identifier - PID for research projects administered by ARDC.
RIMS	Research Information Management System - Example's include Digital Science's Symplectic, Elsevier's PURE, and ReDBox.
RMS	Research Management System - Particularly the ARC RMS that is used for grant applications and reporting.
ROR	Research Organisation Registry - PID for universities, institutions and other research performing organisations.
SHaRED	Submission, Harmonisation and Retrieval of Ecological Data - Research data portal maintained by TERN.
TERN	Terrestrial Ecosystem Research Network - NCRIS infrastructure facility referred to as "Australia's terrestrial ecosystem observatory". Provides data, samples and site-based research infrastructure freely to Australian and international scientific communities.
UA	Universities Australia - Not-for-profit organisation representing universities across Australia, represented by their respective vice-chancellors.
W3ID	World Wide Web Identifier - Stable URL system based on redirects.

Appendix B: Methods and datasets used

This study builds on methods used in previous analyses of the time and opportunity costs involved in research administration. The analysis focuses on the quantifiable time savings that could be achieved with PID-enabled metadata reuse, which would be generated by eliminating the need for manual data entry ('rekeying') to create duplicate records of core research entities.

We gathered data on the number of key entities in the Australian research system, as agreed with the project sponsors at the Australian Access Federation (AAF)[30] and Australian Research Data Commons (ARDC)[31]: researchers, institutions, grants, publications, and projects. Other entities, for example, instruments and geological samples, were of interest to this study, but insufficient data on prevalence and levels of adoption is available at this time.

For publication volumes, we used data from the Dimensions[32] database from Digital Science. The same source was used to estimate the number of research funding grants issued by Australian funders each year, supplemented with data directly obtained from the Australian Research Council[33] and the Medical Research Future Fund[34]. Public data on the number of universities and active researchers were obtained using the latest available Higher Education Research Data Collection (HERDC)[35] dataset. Numbers of projects were derived from estimates given for the number of active projects at any given time in UK universities as part of the UK PID cost-benefit analysis[7], which were adapted to Australia by scaling relative to levels of research funding according to OECD data[36].

We relied on previous work by Research Consulting in the UK, which measured the time taken to input metadata for journal articles, to estimate the time taken for manual data entry, and for the number of data entry events for project metadata[24]. This was found to average 6.73 minutes of staff time per metadata record. In addition, we assumed that for publications, data would have to be rekeyed for each author with an average number of authors per publication of approximately four, based on the work of Fanelli and Larvière[25]. A study undertaken in Norway found that it takes an average of 10 minutes to input basic descriptive information about a research project, and that this data is typically entered into six separate systems[26]. For grants, we assumed that the time burden associated with entering information is similar to that for projects, as they both represent multi-faceted, compound objects with the potential to encompass multiple outputs, people, and

[30] https://aaf.edu.au/
[31] https://ardc.edu.au/
[32] https://www.dimensions.ai/
[33] https://www.arc.gov.au/
[34] https://www.health.gov.au/initiatives-and-programs/medical-research-future-fund
[35] https://www.dese.gov.au/research-block-grants/higher-education-research-data-collection-herdc
[36] https://data.oecd.org/rd/gross-domestic-spending-on-r-d.htm

institutions, so we used the same time estimate of 10 minutes for data entry relating to funding awards.

To estimate the number of manual data entry events for grants and publications, we surveyed Australian institutions on the systems they use for research information management and administration, and on the number of times information about particular entities is manually typed into such systems. Respondents were recruited via existing professional networks, specifically ARMS and CAUL.

Salary costs for three typical roles (a senior researcher or Principle Investigator, a junior researcher, and a research manager) were obtained by averaging data from National Health and Medical Research Council support funding models; job adverts for relevant roles (taken from www.indeed.com and www.au.talent.com); and exemplar universities of varying sizes and intensity of research activity (University of Queensland, MacQuarie University, and Edith Cowan University).

Our savings calculations used an algorithm in which the number of entities of any given type are multiplied by the number of data entry events, multiplied by the time taken in minutes for data entry. This gives us an estimate of the time savings that automated metadata reuse using the records associated with PIDs could bring. These aggregated savings in minutes were converted to days, based on 7.25 working hours per day. We then multiplied the time savings by average salaries to derive financial savings.

Costs for centrally provided services were calculated using budget data shared by ARDC for the full suite of PID services they offer, including Digital Object Identifiers (DOIs) via DataCite, Handles, International Generic Sample Numbers (IGSN), Research Activity identifiers (RAiD), and Persistent Uniform Resource Locators (PURLs) and by AAF for the Australian ORCID consortium, which includes ORCID membership and support project cost recovery.